CSIR DISTINGUISHED LECTURES

BIOMECHANICS

JAMES LIGHTHILL

PUBLICATIONS & INFORMATION DIRECTORATE
Council of Scientific & Industrial Research (CSIR)
New Delhi
1991

Rs 25/-

Sold
4/2/94

BIOMECHANICS

Hearing Sensitivity, Classical Uncertainty and Fish Locomotion

CSIR DISTINGUISHED LECTURES

BIOMECHANICS

Hearing Sensitivity, Classical Uncertainty and Fish Locomotion

James Lighthill
Department of Mathematics
University College London

Publications & Information Directorate
Council of Scientific & Industrial Research
New Delhi

Cover design : Pradip Banerjee
Production : Supriya Gupta

ISBN 81-85038-95-3

© 1991

Publications & Information Directorate
Dr K.S. Krishnan Marg, New Delhi

Designed and Produced by
Publications & Information Directorate
Dr K.S. Krishnan Marg, New Delhi 110 012.

FOREWORD

The main objective of the CSIR Distinguished Lectureship series begun in 1987, is to invite eminent scientists from India and abroad for delivering a series of three lectures on topics of their choice. By the response we have had so far, the scheme had proved to be successful.

This volume contains the three lectures delivered at CSIR, New Delhi; the Central Leather Research Institute, Madras and the National Aeronautical Laboratory, Bangalore, by Sir James Lighthill, London, U.K. As expected, the lectures were well received by scientists of different specializations. In bringing the lectures together in the form of a book, CSIR hopes to make them available to all those who did not have the opportunity of listening to an outstanding scientist with varied interests and achievements.

I express my appreciation of the efforts of CSIR colleagues responsible for the editing and production of this publication.

New Delhi
February 22,1991

A.P. MITRA
Director General, CSIR

CONTENTS

Sl.No.		*Page No.*
1.	Biomechanics of Hearing Sensitivity	1
2.	The Recently Recognised Failure of Predictability in Newtonian Dynamics	43
3.	Biofluiddynamics of Balistiform and Gymnotiform Locomotion	63

BIOMECHANICS OF HEARING SENSITIVITY

Lecture Delivered on 1st December 1989
at Council of Scientific & Industrial Research, New Delhi

PART 1. INTRODUCTION

1. Scope of the Lecture

This biomechanical lecture is concerned with problems of how the reception of sound by human ears, as well as by the ears of many other mammals, is endowed with such astonishing sensitivity. This comprises, not only a highly sensitive frequency discrimination in the ear when viewed as an organ of frequency analysis, but also that extraordinary sensitivity in amplitude discrimination which is represented, at a wide range of frequencies in acoustically healthy human beings, by a 120-decibel range of sound levels between their "threshold of hearing" at the bottom and, at the top, their "threshold of pain" (where intense sounds produce traumatic effects). That's an amplitude ratio of a million to one!

But if I am to be able, within a single lecture, both to tell the exciting story of how scientific knowledge of the **Biomechanics of Hearing Sensitivity** has during the past two decades changed out of all recognition, and also (in telling that story) to explain certain specially important new insights of the last three years, then I must carefully concentrate my narrative. First, I must confine it to the biomechanics of hearing sensitivity in man, cat and guineapig, and in the handful of other primate, carnivore and rodent species where the experiments most crucial to the advance of the subject have been made; consequently, many fascinating issues (for example, about hearing in bats or dolphins) will be completely ignored. Next, I shall limit my attention to the reception and processing of acoustic signals within the inner ear (Fig.1) and its associated nervous system; thus, the middle ear's transmission mechanism, a subtle mechanical device whereby sound energy channelled through the air within the outer ear and impinging on the ear drum is conveyed with very little loss (once again, in acoustically healthy subjects) to the aqueous medium of the inner ear via its oval window (Fig.1), must be put on one side — just from shortage of time. More generally, I shall be obliged to simplify all explanations to quite a drastic extent; while, however, offering in my text some detailed references to much fuller treatments of each particular topic.

2. Cochlear anatomy: an elementary account

That part of the mammalian inner ear which responds to acoustic signals (rather than to accelerations or rotations of the head as do the semicircular canals) is the spirally coiled organ known as the cochlea (Fig.1). Embedded within the tympanic bone (which gives it excellent

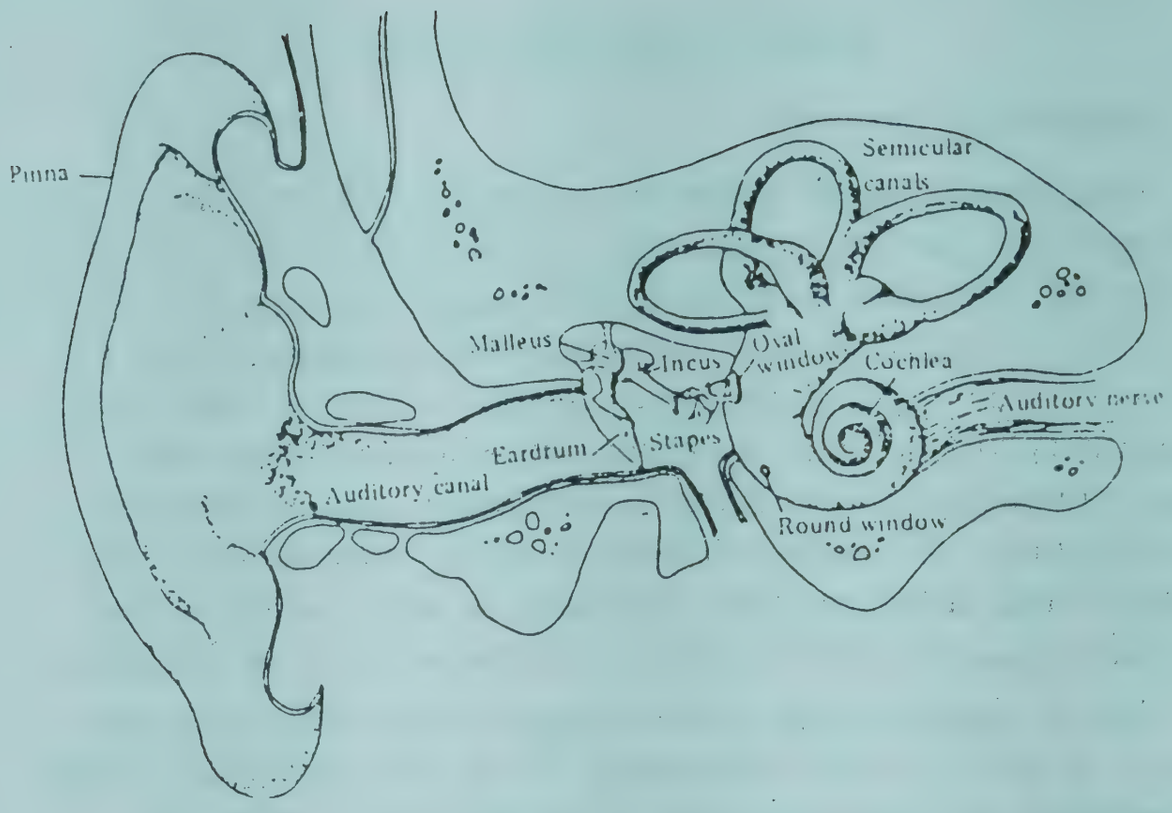


Fig.1. Schematic drawing of the human ear (reproduced, with permission, from Lindsay & Norman 1977, p. 126). Note the middle ear's linkage mechanism, which transfers acoustic energy from vibrations of the eardrum to vibrations of the oval window.

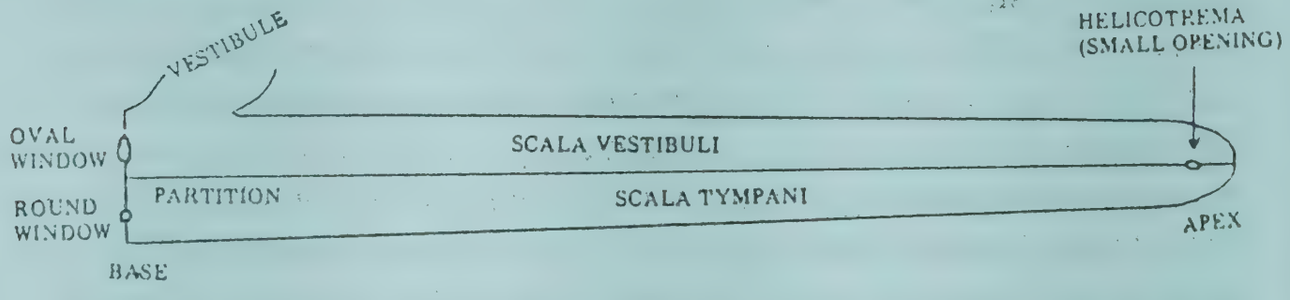


Fig.2. Diagram of the cochlea from a broad macroscopic viewpoint, with its spirally curved centreline (of length around 35mm) shown for convenience as stretched straight. A partition divides the cochlea into two tubes (scala vestibuli and scala tympani), each of cross-section around 1 mm^2 and filled with an aqueous medium: perilymph. The middle-ear linkage mechanism applies forcing to the oval window (in scala vestibuli) while the round window (in scala tympani) is free. Note: The macromechanical analysis that follows in Part 2 takes precisely the same form whether the spiral shape of the cochlea's centreline is taken into account or, as in this diagram, ignored. Essentially, the function of the spiral shape seems to be to allow a relatively long cochlea to be incorporated within a compact tympanic bone.

protection from mechanical damage), its coiled passage has an internal diameter of about one and a half millimetres in human subjects and a total length of about 35mm measured along its spirally curved centreline. From a broad macroscopic point of view (Fig.2), a partition extending along this entire length of the cochlea divides it into two chambers: the scala vestibuli and scala tympani, each of cross-section about 1 mm^2 and filled

with an aqueous medium called perilymph. The partition dividing them has as we shall see a most complex and important microscopic structure; however, from the macroscopic standpoint, its principal mechanical property is a continuous variation of stiffness (by close on four orders of magnitude) along its length.

We shall find that other important mechanical properties of the cochlear partition in living animals have become known only during the past ten years. But the fact that its stiffness is enormously greater near the base (that is, at the point of entry of the acoustic signal) than near the apex has long been recognized as implying characteristic frequencies of vibration at cochlear cross-sections which vary in man from a high value (perhaps 20 kHz) near the base to a low value (perhaps 200 Hz) near the apex. This gives a first insight into the frequency analysing powers of the ear: across that wide range of frequencies acoustic components of every different frequency find their way to a different characteristic place along the cochlea (by means which we shall discuss) with higher frequencies reaching a characteristic place which is near the base and lower frequencies a place relatively closer to the apex, in a continuously graded sequence.

Going now beyond the macroscopic scale. I describe the cochlear partition first of all at an intermediate level of resolution. On this scale it is essentially an internal duct, the scala media (Fig.3), filled with another aqueous medium called endolymph. From a purely mechanical point of view, perilymph and endolymph have practically identical properties (i.e. those of water); but endolymph, continuously secreted within the scala media by the stria vascularis, has a quite different chemical constitution that is essential to the maintenance of the sensitive hair cells whose response to the acoustic signal plays a critical role in hearing.

The scala media, with a cross-sectional area of about $0.2mm^2$, is separated from the scala vestibuli by Reissner's membrane and from the scala tympani by the basilar membrane. Reissner's membrane is extremely compliant, being just three cells thick, and can be regarded as simply moving with the surrounding fluid.

Essentially, then, the stiffness of the cochlear partition resides in the basilar membrane, stretched across part of the cochlear diameter between the bony shelf on the left of Fig.3 and the (also essentially rigid) spiral ligament on the right. The partition's stiffness, in the sense of the pressure rise in the scala tympani required to make a given increase in its area by bending this basilar membrane, becomes less and less (as already stated)

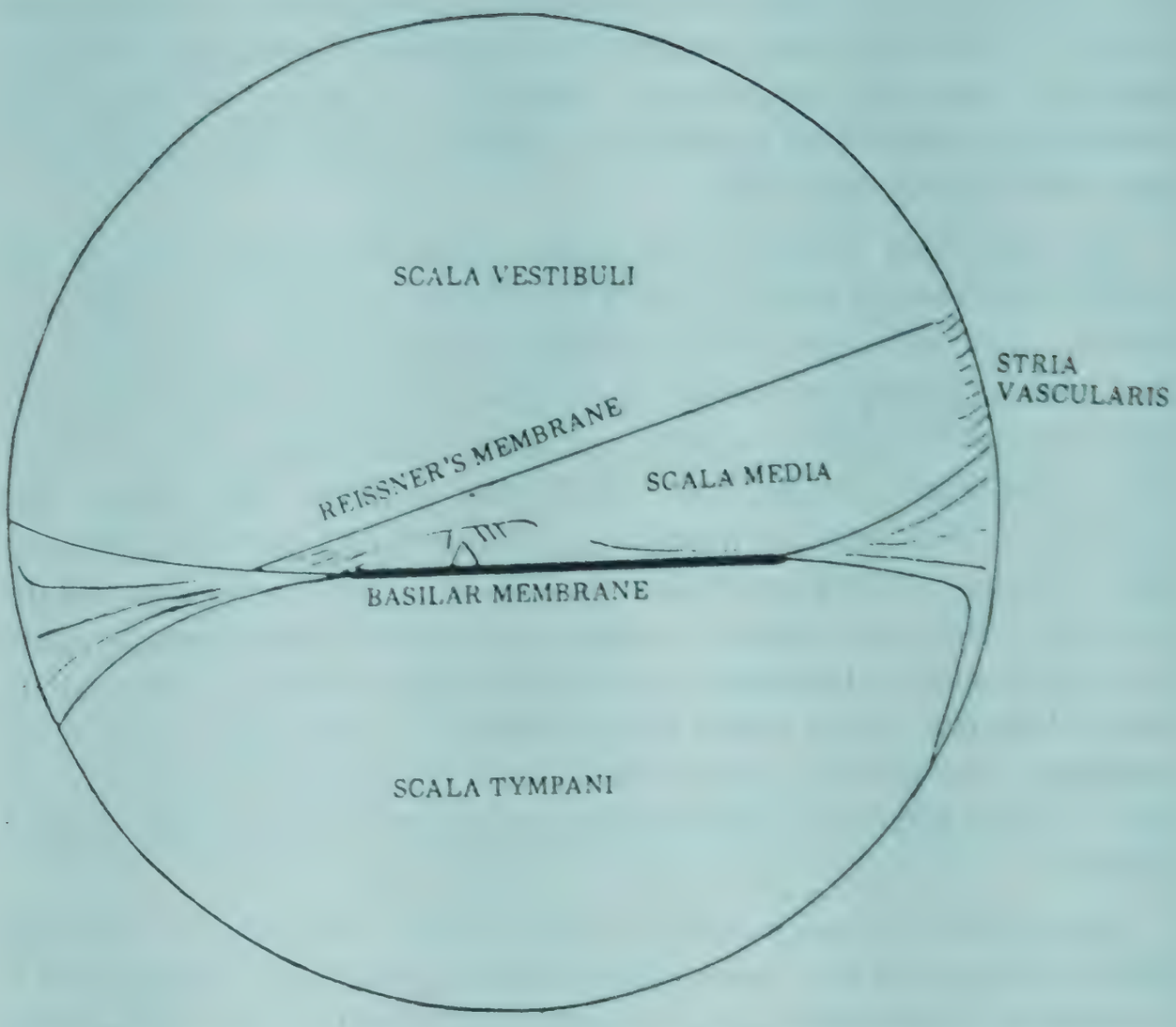


Fig.3. Diagram of a cochlear cross-section at an intermediate level of resolution, revealing the cochlear partition as an internal duct: the scala media, of cross-section around $0.2mm^2$, filled with a different aqueous medium (endolymph). The scala media is separated from the scala vestibuli by the very flimsy Reissner's membrane, and from the scala tympani by the basilar membrane. Essentially, all the stiffness of the cochlear partition resides in the basilar membrane, stretched between the bony shelf (on the left) and the spiral ligament (on the right). The massive increase in its compliance (decrease in stiffness) from base to apex is due primarily to a gradual increase in width (from 0.1mm to 0.5mm) and a corresponding decrease in thickness.

by a large factor from base to apex. This results primarily from a gradual increase in width of the basilar membrane (from about 0.1mm to 0.5mm) and a corresponding decrease in thickness; both of which trends contribute to the decrease in stiffness.

Proceeding now to a microscopic scale at the level of single cells within the so-called organ of Corti (Fig.4) I want to draw special attention to the presence at each cross-section of the cochlea of four so-called hair cells: one inner hair cell and three outer hair cells. The name "hair cell" refers to the fact that each cell tip has a "hairy" appearance resulting from the

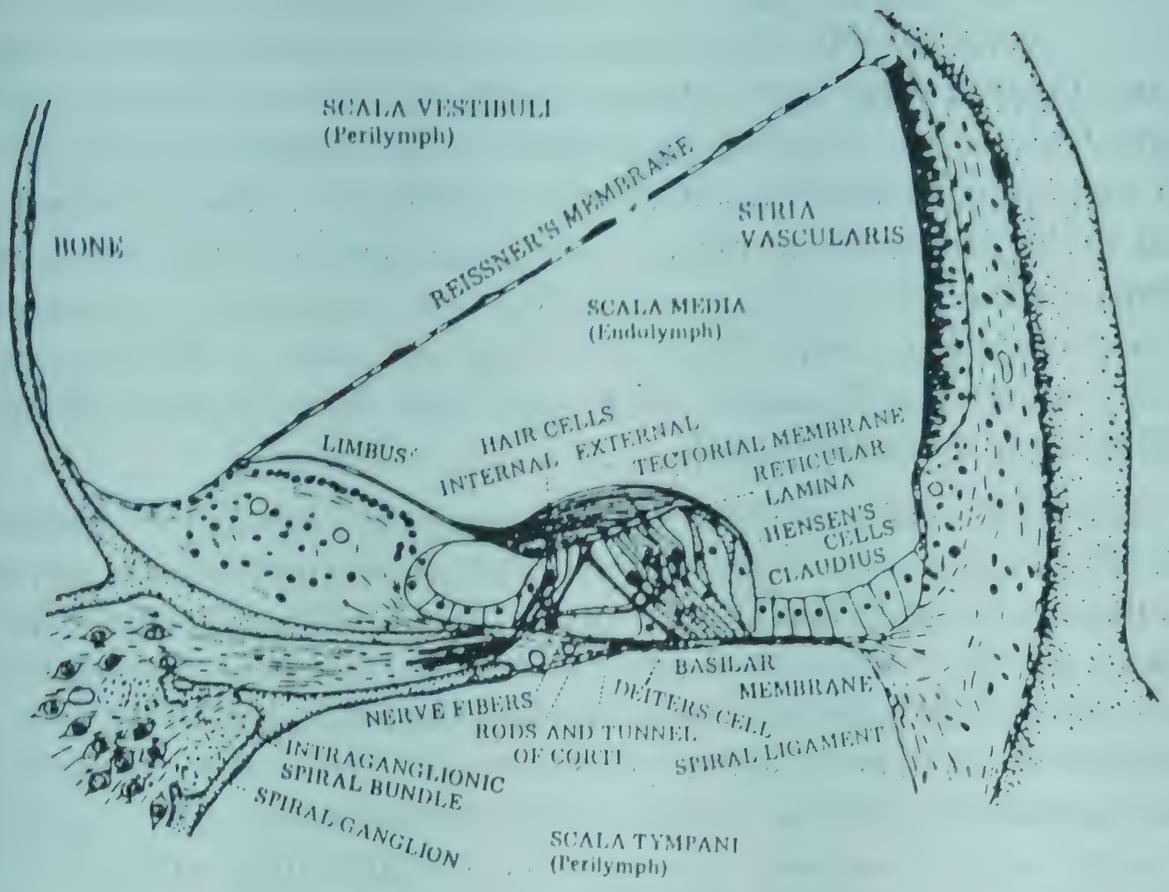


Fig.4. Representation of a cross-section of a guineapig's cochlea on a (high-resolution) microscopic scale, at the level of single cells within the organ of Corti (Davis *et al* 1953). Note especially one "internal" or *inner hair cell* and three "external" or *outer hair cells*.

presence of highly specialised stereocilia, arranged in a characteristic pattern. These stereocilia respond very sensitively to even a weak acoustic signal. Typically, this is a signal which has the frequency required to find its way to that characteristic place along the length of the cochlea.

All hair cells have nerve fibres attached to them. For inner hair cells they are afferent fibres; in other words, nerve fibres that carry signals from the cell towards the brain. For this reason it was long suspected (Békésy, 1960) that the inner hair cells were the transducers which convert an acoustic signal into a neural signal (travelling along a nerve fibre), and all recent evidence has fully confirmed that belief. The exact mechanism of the transduction process in a mammalian cochlea[1] remains uncertain. It must be influenced, however, by the way in which the inner hair cell is suspended (Fig.4) between the basilar membrane and the much flimsier tectorial membrane.

[1]Where the interesting transduction mechanism discovered by Crawford & Fettiplace (1980) in turtle hair cells is, perhaps unlikely to be present (Hudspeth, 1985).

In a human cochlea, some 4,000 inner hair cells are stretched along its length, and each sends a signal to the brain along a group of nerve fibres, which together form part of the complete auditory nerve. Each of these nerve fibres comes from a particular place along the cochlea and is, above all, carrying information about acoustic signals at frequencies for which that is the characteristic place. Indeed, electrophysiological measurements of the neural activity in one of those fibres show a level of response to acoustic signals which has a very sharp maximum at that particular frequency (Kiang, Watanabe, Thomas & Clark, 1965; Evans & Wilson, 1973; Kiang & Moxon, 1974).

In addition, there are three times as many outer hair cells: 12,000 in all. These outer hair cells are endowed with efferent nerve fibres carrying to them, *from* more central parts of the nervous system, neural signals which can be presumed to control their function. The long-debated question of what is the function of the outer hair cells seems at last to have been answered in the past three years; however, in this first very elementary account of cochlear anatomy I shall not anticipate either that discovery or the brilliant research work (including work on the sub-microscopic anatomy of hair cells) that has led to it.

3. Cochlear sensitivity: a brief initial outline

The remainder of this biomechanical lecture in fact, is divided into two main parts. The first of these (part 2) which I call macromechanical, describes from the broad macroscopic viewpoint (Fig.2) combined with the intermediate level of resolution (Fig.3) our experimental and theoretical knowledge of those vibratory movements of the basilar membrane which themselves possess the remarkable property already mentioned; namely, that different frequency components of an acoustic signal "find their way", as I put it, to different characteristic places along the cochlea. And the subsequent part (Part 3) is micromechanical, being concerned at the microscopic scale (Fig.4) with all aspects of how, at such a characteristic place, the organ of Corti as a whole responds to the local vibratory movements of the basilar membrane within the ambient fluid medium.

But before I embark in detail on either the macromechanics or the micromechanics of the cochlea, I want to offer this distinguished but diverse audience, with its great variety of different backgrounds, some help in following what for most of you may be relatively unfamiliar subject matter. I am including at this point, therefore, a brief initial outline of both the macromechanical and the micromechanical material, that will be treated more fully in the lecture's later parts.

Now many of you, I believe, may be familiar with the biomechanics of flow in a distensible tube; for example, blood flow in an artery. The greater the tube's distensibility D (defined as the proportional increase in cross-sectional area resulting from a unit rise in internal pressure) the slower is the resulting propagation of fluid motions along the tube, at a pulse speed that takes, actually, the form[2]

$$(\rho D)^{-1/2} \qquad \text{(where } \rho \text{ is fluid density)} \qquad \ldots (1)$$

varying inversely with the square root of the distensibility D.

The cochlea however is not a single distensible tube but comprises a pair of adjacent tubes (Fig.2), embedded in the practically rigid tympanic bone and separated by a compliant partition; of which, furthermore, the compliance increases from base to apex. It is interesting to identify the types of wave motion possible in such a pair of adjacent tubes, and then to consider the effect of forcing motions applied at the base to one of the tubes (the scala vestibuli, forced at the oval window) while motions at the base of the other tube (the scala tympani, unconstrained at the round window) remain completely free.

Actually,any wave motion in the two tubes can be thought of (Peterson & Bogert, 1950; see also Lighthill, 1981) as a combination of:

(i) a wave with *equal* pressure changes in the two tubes, and therefore no movement in the partition - this is a fast-moving wave since its speed (1) depends on the extremely low distensibility of the surrounding bony tissue; and

(ii) a wave with equal and *opposite* pressure changes in the tubes, and therefore with their difference in pressure producing movement of the partition - this is a relatively slow-moving wave whose speed (furthermore), being determined rather as in (1) by the compliance of the partition[3] tends to decrease steeply with distance from the base as that compliance increases.

---

[2]A more accurate form (Lighthill, 1978b, p.93) for the wave speed can be written $[\rho(K+D)]^{-1/2}$ in terms of the sum of the fluid's compressibility K and the tube's distensibility D; in biomechanical applications, however, K is small compared with D.

[3]Indeed, if an "equivalent distensibility" D is defined as the proportional increase in the cross-sectional area of one tube from a unit pressure rise in *that* tube (accompanied by an equal pressure fall in the other), then the same equations are satisfied — at least for "long waves" — as in a single distensible tube. Specifically, this definition gives $D\,\partial p/\partial t = A^{-1}\partial A/\partial t = -A^{-1}\partial J/\partial x$ by the equation of continuity (where J is volume flow in the tube) but the rate of change of momentum per unit length, $\rho\partial J/\partial t$ is equal to a pressure force $-A\partial p/\partial x$, which allows J to be eliminated from the previous equation to give $\rho D\partial^2 p/\partial t^2 = \partial^2 p/\partial x^2$, representing waves of speed $(\rho D)^{-1/2}$.

If forcing is applied in one tube only at the base, where the other tube remains free, then the fast and slow waves must be generated with equal amplitudes of pressure change at the base itself; so that, although reinforcing each other in the scala vestibuli, they cancel out (at the base, again) in the scala tympani. However, with distance from the base, the slow wave progressively travels more and more slowly, while it grows in amplitude and so becomes the dominant component; the fast wave, by contrast, adopts the simple standing-wave pattern which in acoustics we call an organ-pipe mode and exhibits no amplitude increase, being also uninteresting in another way as producing no motion of the cochlear partition.

Accordingly, the fast wave becomes quite unimportant and I shall omit any further mention of it [4]. The only wave referred to hereafter will be the slow wave with equal and opposite pressure changes on both sides of the partition.

Now this slow wave shares some important properties with yet another famous type of wave motion from classical fluid mechanics: namely, water waves propagated along an open channel. And one of these properties - the one we call dispersion - is present for much the same reason in the two cases. So I shall briefly remind you about such waves in open channels.

Really an open channel is very like a distensible tube (Fig.5), at least for the propagation of long waves where the pressure change is the same over the entire cross-section. Then, when the free surface is raised a distance δ, this pressure increases by $\rho g \delta$ whilst the cross-sectional area A of the channel increases by $b\delta$ (where b is the free-surface breadth); effectively, then, the distensibility, relating the proportional increase $b\delta/A$ in cross-sectional area to the pressure increase $\rho g \delta$, is

$$D = b/A\rho g \qquad \ldots (2)$$

and the speed (1) of long waves in the channel takes the classical form

$$(gA/b)^{1/2} \qquad \ldots (3)$$

in terms of the "hydraulic mean depth" A/b.

We know however that, as the wavelength becomes shorter, there ceases to be any uniformity in the distribution of pressure changes and of consequential fluid motions over the cross-section. Rather, they become

---

[4] Except in a footnote (no.11) in section 7.

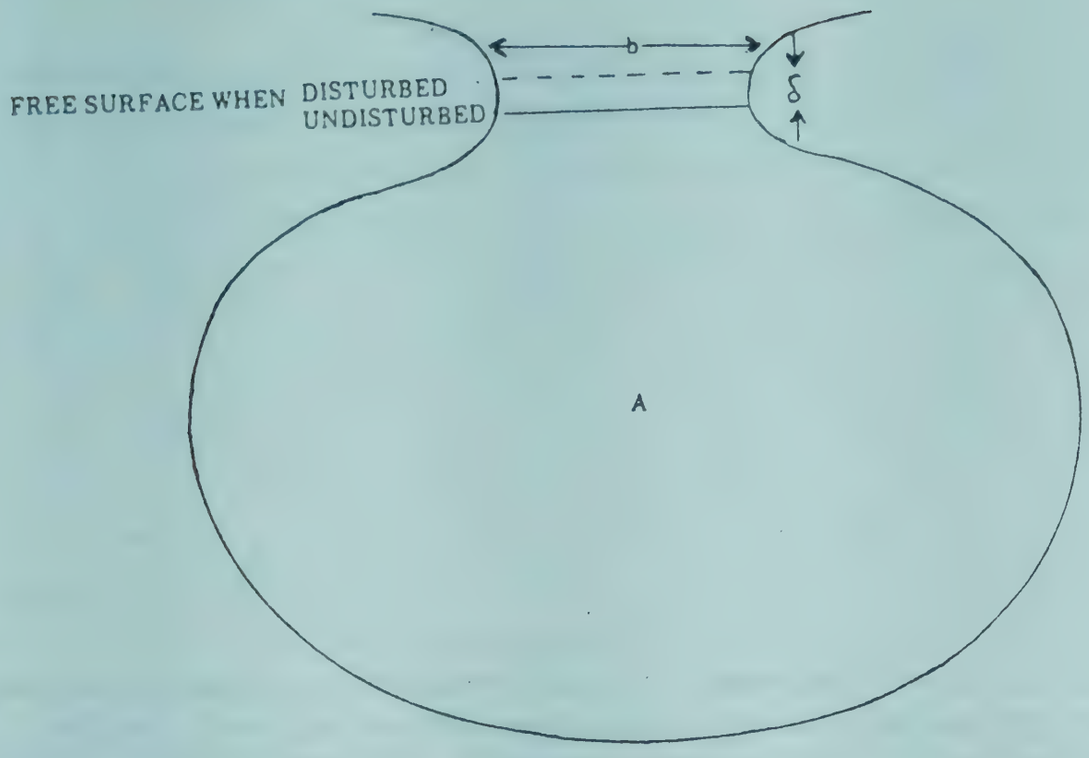


Fig.5. Waves in open channels (the diagram shows the channel cross-section, taken as uniform along the length of the channel). Case (i): propagation of long waves, with fluid motions filling the channel. When the free surface is raised a distance δ, the pressure is increased by $\rho g\delta$ (where ρ is the fluid density) while the cross-sectional area A is increased by $b\delta$ (where b is the breadth of the free surface); the channel can therefore be said to have distensibility $b/A\rho g$.

more and more concentrated near the free surface, within a distance from it of about a wavelength divided by 2π, as shown in Fig.6(a). This causes the wave velocity to fall below the long-wave form (3); and, in fact, the so-called phase velocity with which the crests travel becomes less than the long-wave speed *in a manner depending on wavelength* (this is the dispersive property). It is also important that the group velocity, which specifies how fast the energy in a group of waves travels (how fast a signal travels, if you like) falls even more than the phase velocity itself for these dispersive waves.

A similar cause of dispersion exists in the cochlea's two adjacent tubes, each of which is distensible at just one part of its boundary: the basilar membrane, playing a role like that of the free surface for an open channel. As the wavelength becomes less, the fluid motions in both tubes become concentrated more and more around the membrane, as shown in Fig.6(b). This reduces the phase velocity of the wave, in a manner depending on wavelength; and therefore reduces the group velocity even more, slowing down the signal as it travels along the basilar membrane.

So in these remarks I have prepared you for the important role played by wave dispersion in the cochlea. Actually, one further cause of disper-

Fig.6. (a) Waves in open channels. Case (ii): propagation of short waves (of wavelength λ), with fluid motions penetrating a mean distance $\lambda/2\pi$ into the channel. The paths of fluid particles are roughly circular, in planes perpendicular to the paper, with radii greatest near the free surface. Contours of constant kinetic energy are sketched as broken lines (it is the logarithm of the energy which changes by equal amounts between adjacent contours).

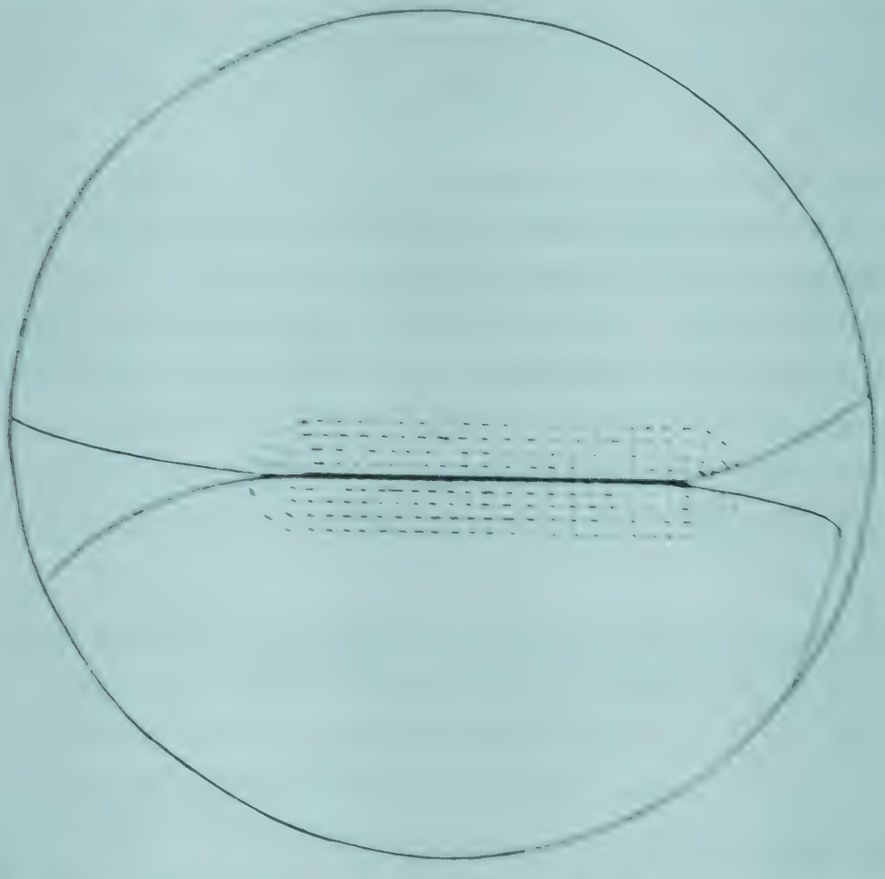


Fig.6 (b) Corresponding diagram for short waves in the cochlea. Each of its two fluid-filled tubes is distensible at just one part of its boundary (the basilar membrane, analogous to the channel's free surface) so that, in both tubes, motions associated with shorter waves are concentrated more and more around the membrane. The broken lines again represent contours of constant kinetic energy.

sion exists, as we shall see, in the form of the inertia of the cochlear partition. The two causes taken together produce very pronounced dispersion, so that the group velocity falls sharply, decreasing all the way to zero — whether we consider it at a fixed location as wavelength decreases or at a fixed frequency as waves of that frequency travel along the cochlea. This means that energy of a given frequency progressively slows down until it comes to rest at the characteristic place for that frequency. Here it "piles up" in the form of a local oscillation which is damped only by the effects of fluid viscosity, acting within thin Stokes boundary layers.

The importance of this mechanical contribution to cochlear sensitivity became clearer and clearer in the past twenty years as accurate measurements of basilar-membrane vibrations in living animals were made, using new techniques of *in vivo* observation. The so-called Mössbauer source emits gamma rays of an exactly determined frequency, so that for such a source velocities of vibration as low as a few micrometres per second can be determined from the Doppler shifts which they impart to that frequency of gamma-ray emissions. Gifted experimenters were able to implant a Mössbauer source on the basilar membrane, gaining access to it from the scala tympani (Johnstone, Taylor & Boyle, 1970; Rhode, 1971). From the outset their results demonstrated a rather sharp tuning of the basilar-membrane vibrations at the point where the Mössbauer source is implanted in response to acoustic signals of varying frequency.

Then, with increasing refinement of the techniques for making the measurements under conditions of minimum trauma (Sellick, Patuzzi & Johnstone, 1982; Robles, Ruggero & Rich, 1986) it became clear that the basilar-membrane response has a very sharp maximum indeed at the frequency for which that point is the characteristic place. The also very sharp tuning of auditory nerve fibres which I mentioned earlier (that is, their sharp maximum of response to acoustic signals at a characteristic frequency for each fibre) is found in these recent measurements to be fully matched by the sharp tuning of the basilar-membrane response (Fig.7). This is what would be expected if the transduction mechanism in the inner hair cells simply converts basilar membrane vibration into neural activity.

In the meantime, improved measurements of the mechanical properties of the basilar membrane in living animals (Voldrich, 1978) had demonstrated how neighbouring short sections of the membrane vibrate almost independently of one another. This highly anisotropic property of the living membrane, which had not previously been recognized (since in cadavers the membrane degenerates into something like an isotropic

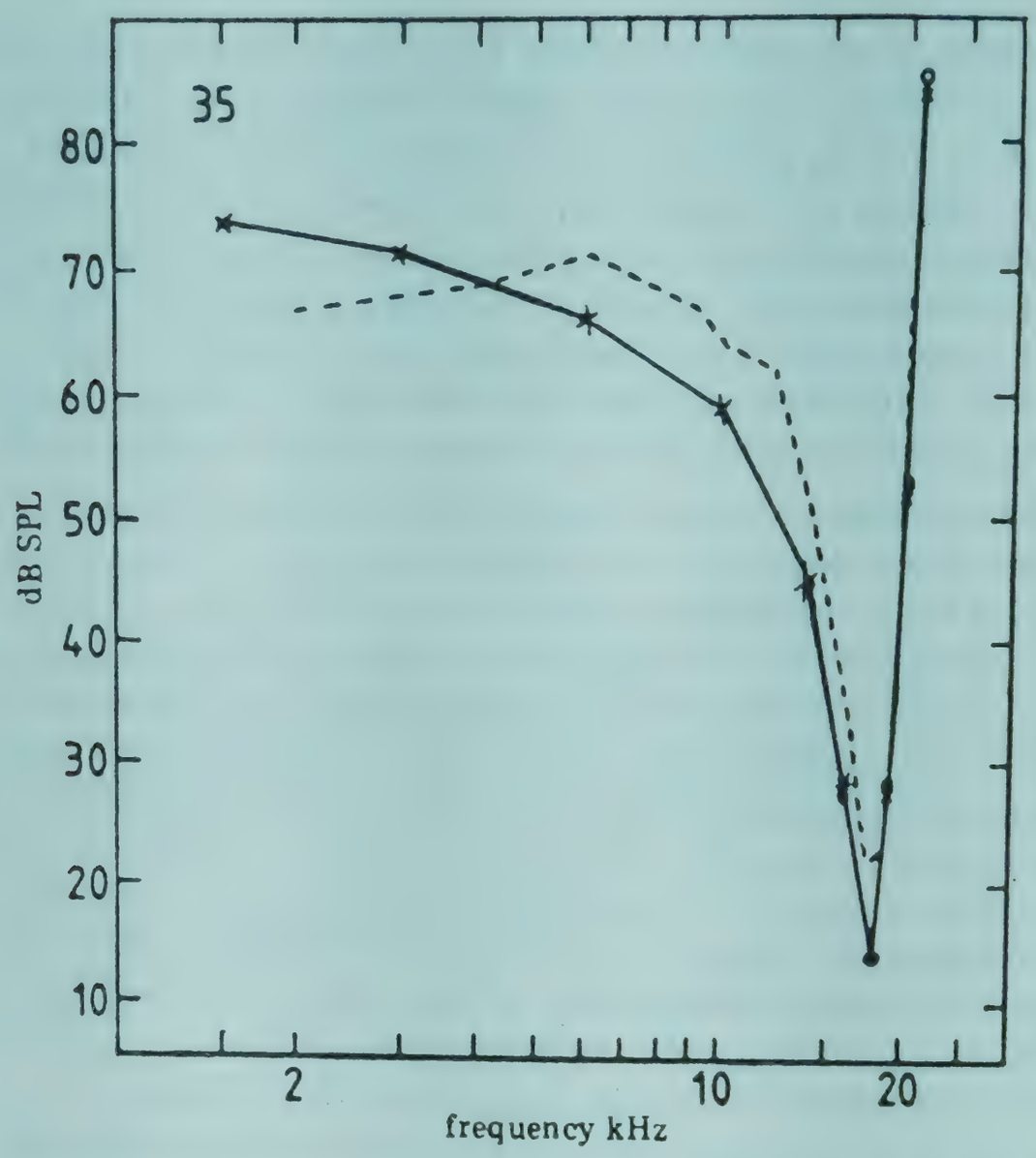


Fig.7. A pair of comparable threshold curves, as a function of frequency, for the guineapig cochlea (Sellick, Patuzzi & Johnstone, 1982). [*Plain line*: sound pressure level (in dB) required to generate (at a particular place where a Mössbauer source was implanted) vibrations of the basilar membrane with velocity amplitude 0.04mm/s. *Broken Line*: Sound pressure level (in dB) required to generate activity in a particular auditory nerve fibre (emanating from essentially the same place)]. In both curves the frequency for which the threshold is extremely low is that frequency for which the characteristic place was where the source was implanted.

elastic plate), facilitates the dispersive behaviour which I mentioned earlier. Essentially, it permits the travelling wave at a given frequency to undergo a progressive reduction in wavelength, and therefore also in group velocity, to zero as the characteristic place is approached. Then its amplitude must build up in a way limited only by viscous dissipation, which allows a very sharp tuning in the basilar-membrane vibrations.

Nevertheless, even that modest level of energy dissipation by viscous action is now believed to prevent this remarkable mechanism of frequency

selectivity based on the passive mechanical properties of the basilar membrane from accounting entirely for the extreme sharpness of tuning observed in the best recent measurements. Later on, I shall be outlining compelling evidence to the effect that an active feedback mechanism which resides in the outer hair cells produces a further sharpening-up of the basilar membrane's response at low amplitudes, and so makes a vital contribution to hearing sensitivity.

PART 2. COCHLEAR MACROMECHANICS

4. Experimental evidence for dispersion

But first of all I shall give a more systematic account of the macromechanics of the cochlea, relating as far as possible our experimental and theoretical knowledge of the vibrations of the basilar membrane, not only from the broad macroscopic point of view (Fig.2) but also at the intermediate level of resolution (Fig.3). At this latter level we have to think of vibrations of the basilar membrane as generating fluid motions in two adjacent tubes (as before); out of which, however, while one is simply the scala tympani, the other must be regarded as a combination of the scala vestibuli and the scala media, separated by the very flimsy Reissner's membrane (just three cells thick). The perilymph and endolymph, although chemically distinct fluids, both share practically the same *mechanical* properties (density and viscosity, that is) as those of water, while Reissner's membrane (again of the same density) must respond to pressure distributions in much the same way as does the surrounding fluid. Therefore the fluid mechanics should be essentially the same in each of the two tubes separated by the basilar membrane.

And all of that is important in the context of the dispersive wave I described to you; one in which, for a given frequency, the wavelength gets less and less as the characteristic place is approached. Then, as I foreshadowed by reminding you of the properties of short waves in an open channel (Fig.6), the fluid motions are expected to be concentrated very close to the membrane, at a distance of order the wavelength divided by 2π. This means that, rather near the characteristic place, these fluid motions are likely to be much less energetic in the scala vestibuli proper than in the scala media; which, of course, is just where they can most strongly influence the organ of Corti (see Part 3).

But before I go any further, I want to offer you a clear experimental proof that waves propagated along the basilar membrane are indeed

dispersive. For this purpose I will use the excellent data of William Rhode of the University of Wisconsin (Rhode, 1971), obtained by the Mössbauer-source technique in live squirrel monkeys. He used a surgical procedure allowing a small Mössbauer source of diameter 0.06mm to be implanted on the basilar membrane, after which the squirrel monkey and, in particular, its cochlea could be restored to a normal condition (with, for example, replacement of any fluid lost from the scala tympani, and with the access hole into it through the bone securely plugged up). Rhode had also placed another Mössbauer source on the point of the eardrum which is in direct contact with the handle of the malleus bone in the middle ear; it is this malleus bone whose movements, via the linkage mechanism in the middle ear (Fig.1), drive the motions of the oval window.

Thus, when the ear was stimulated acoustically by a pure tone, the sinusoidal vibrations *both* of the basilar membrane *and* of the malleus - which drives it - could be determined in both amplitude and phase (such values being derived, you will remember, from the time-variation of the Doppler-shifted frequencies of those gamma rays which the sources emit). It is the amplitude *ratios* and phase *differences* that are significant. Thus, if the displacements of the two sources are

$$a_M \cos(\omega t + \Theta_M) \text{ and } a_B \cos(\omega t + \Theta_B) \quad \ldots (4)$$

for the malleus and basilar membrane, Rhode plots the amplitude ratio a_B/a_M and phase difference $\Theta_B - \Theta_M$ as functions of the frequency in hertz, $\omega/2\pi$. The amplitude ratio is shown in Fig.8 on a decibel scale (thus, the quantity plotted is $20 \log_{10}(a_B/a_M)$ against frequency on a logarithmic scale, while the phase difference is plotted in radians in Fig.9 against frequency on a linear scale.

The asymmetric peak in Fig.8 reflects, of course, the basic properties of the travelling-wave system on the basilar membrane. Actually, the Mössbauer source in this case was placed fairly near the base so that it was acoustic signals of a relatively high frequency, 7kHz, which excited maximum response; this location, in short, was the characteristic place towards which such signals of frequency 7kHz travel and there "pile up" into a local oscillation (limited in magnitude by viscous dissipation). Any signals at a higher frequency than that, however, will have "piled up" (and, ultimately, become dissipated) at a point still nearer to the base, so as not to reach this Mössbauer source at all. By contrast, signals at a lower frequency would as they passed the source have already substantially increased in amplitude on their way to a characteristic place nearer to the

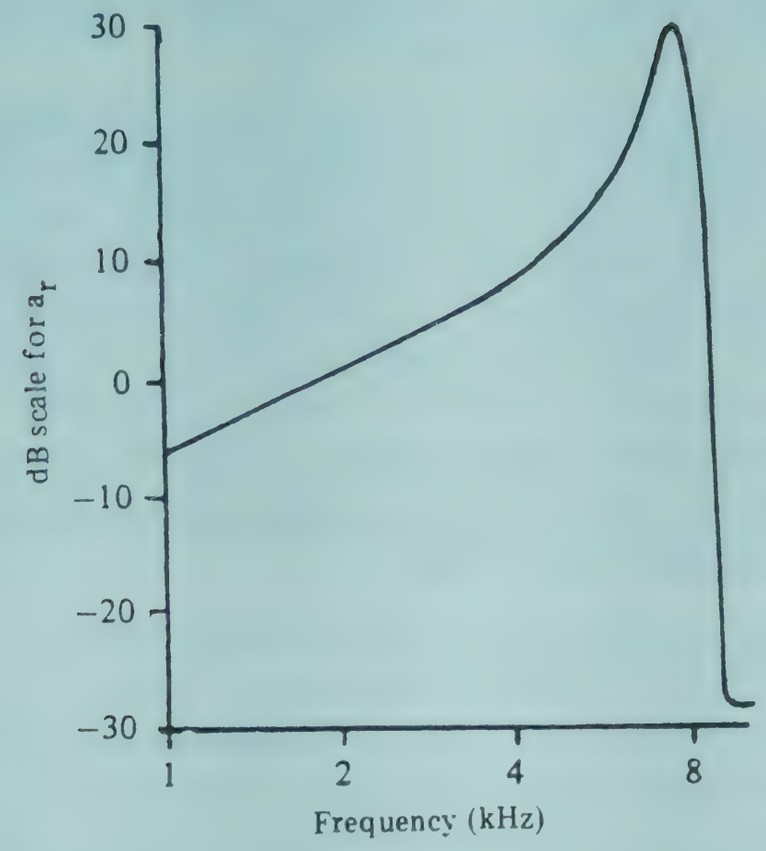


Fig.8. Amplitude ratio $a_B / a_M = a_r$ shown on a decibel scale (thus, the quantity plotted is $20 \log_{10} a_r$) as a function of frequency (in kHz), also shown on a logarithmic scale: data obtained by Rhode (1971) for animal 69.473, with sound intensity 70dB used near resonant frequency. The slope of the curve is 6dB/octave at frequencies below about 5kHz, while being markedly different at higher frequencies.

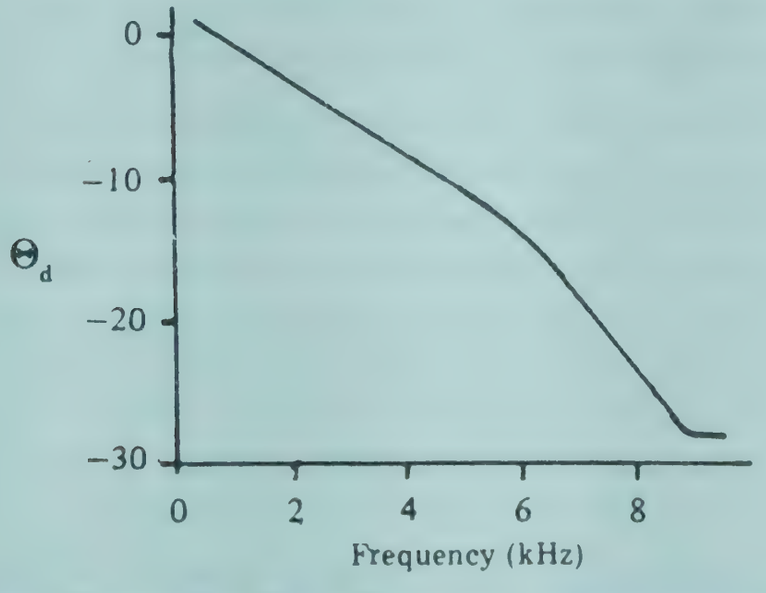


Fig.9. Phase difference $\Theta_B - \Theta_M = \Theta_d$ plotted in radians as a function of frequency in kHz by Rhode (1971) for animal 69.662, with sound intensity 85dB used near resonant frequency. The curve satisfies equation (5) at frequencies below about 5kHz but deviates significantly from it at higher frequencies.

apex. The peak, then, is very steep on the high-frequency side but relatively gradual on the low-frequency side.

Now I come to the evidence for dispersion. If the travelling wave were nondispersive, in other words if at each point of the basilar membrane the wave had a definite velocity — greater near the base and less near the apex, of course, but without any dependence on frequency, then there would be a definite travel time τ for waves of every frequency between the base and the location of the Mössbauer source. Then the phase for vibrations of the basilar membrane would take the form Θ_B = constant – $\omega\tau$ in terms of the radian frequency ω and this fixed travel time. And by considering the low-frequency limit of this equation we can determine the constant so as to obtain

$$\Theta_B = \Theta_M + (1/2)\pi - \omega\tau \qquad \ldots (5)$$

This determination makes use of the fact that the low-frequency limit is a stiffness-dominated regime where basilar-membrane displacement is in phase with the (equal and opposite) pressures on its two sides; and that these pressures, just as within a wave in a single distensible tube (or for that matter in a sound wave), are in phase with the longitudinal velocity of the fluid. Thus, in the low-frequency limit where Θ_B becomes effectively the phase of pressure fluctuations at the base, it must be that of the oval window's velocity and hence also of the malleus velocity, so it is ahead of the phase of the malleus displacement by $(1/2)\pi$. The same dependence of basilar-membrane displacement on malleus velocity in the nondispersive case would lead us to expect a constant slope (of 6dB/octave) for the amplitude-ratio against frequency in the log-log plot of Fig.8.

Now, Rhode's data confirm that both these expectations for nondispersive waves - equation (5) for phase and the 6dB/octave slope for amplitude - are fulfilled for signals at frequencies below about 5kHz, which still have a lot farther to go before they reach their characteristic place. At higher frequencies, on the other hand, there are big deviations, which have to be attributed to dispersion. Often, we try to analyse graphically the nature of dispersion by means of a plot of the radian frequency ω against the quantity k which we call the wavenumber. In any travelling wave, the wavenumber k is defined as the rate of reduction of phase with distance of travel. For a sine-wave, of course, k is just 2π divided by the wavelength; but evidently we are *not* dealing with sinusoidal waves in a system with steep gradients of wave velocity like the basilar membrane. Nevertheless, another of Rhode's experiments allowed a direct estimate

of the wavenumber k as the (downward) gradient of phase along the membrane.

This was the experiment (see Fig.10) where the phase difference $\Theta_B - \Theta_M$ was measured simultaneously at two Mössbauer sources (1 and 2, each of size 0.06mm) which had been implanted at different locations on the basilar membrane, 1.5mm apart. The definition of the wavenumber k implies that its *average value* k_{av} in the region between the two sources is given by the divided difference

$$k_{av} = (\Theta_{B1} - \Theta_{B2}) / (1.5\text{mm}) \qquad \ldots (6)$$

Values of k_{av} are plotted as abscissa in Fig.11 with the frequency ω as ordinate.

In such a plot of frequency versus wavenumber we expect for nondispersive waves a straight line through the origin at a slope equal to the wave velocity. Such nondispersive behaviour is found as before at frequencies below about 5kHz and the associated wave velocity for all these frequencies is 15 m/s. At higher frequencies, however, the highly dispersive character of the waves is shown by the fact that the curve as far as it goes

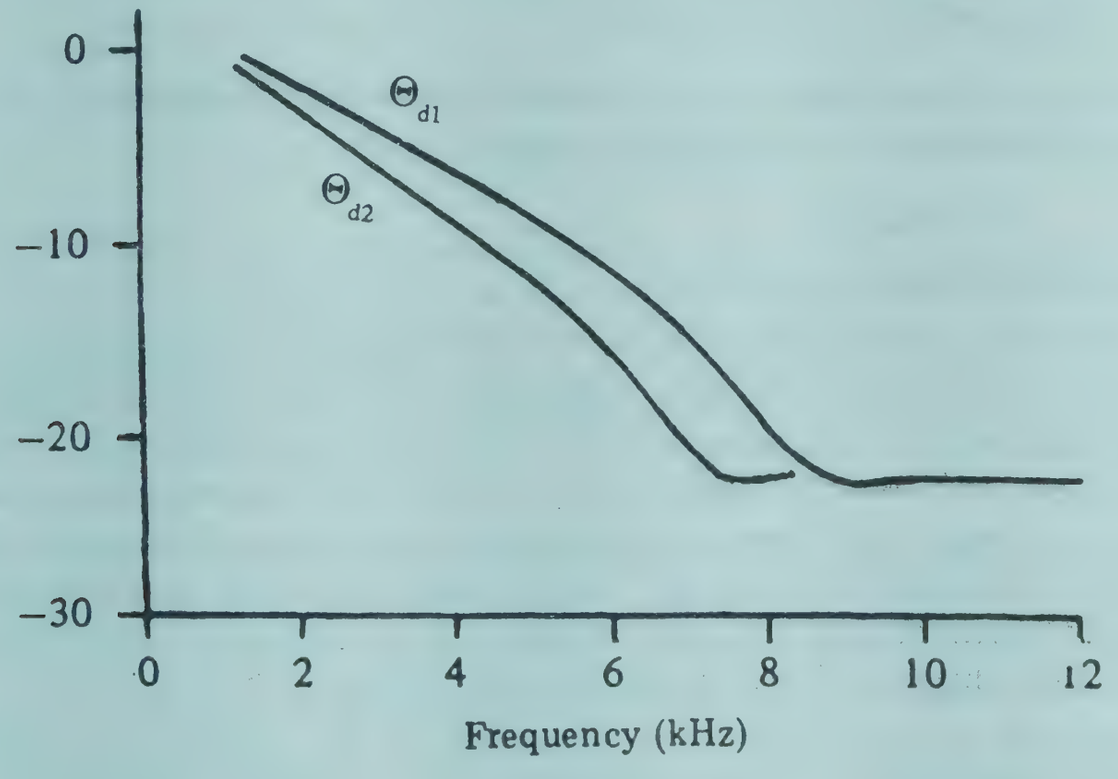


Fig.10. Phase differences Θ_{d1} and Θ_{d2} obtained (Rhode, 1971) as in figure 9 at two different points 1 and 2 on the basilar membrane (1.5mm apart) in animal 69.435. The source size (0.06mm) was demonstrated by Rhode to be too small to influence the validity of the data.

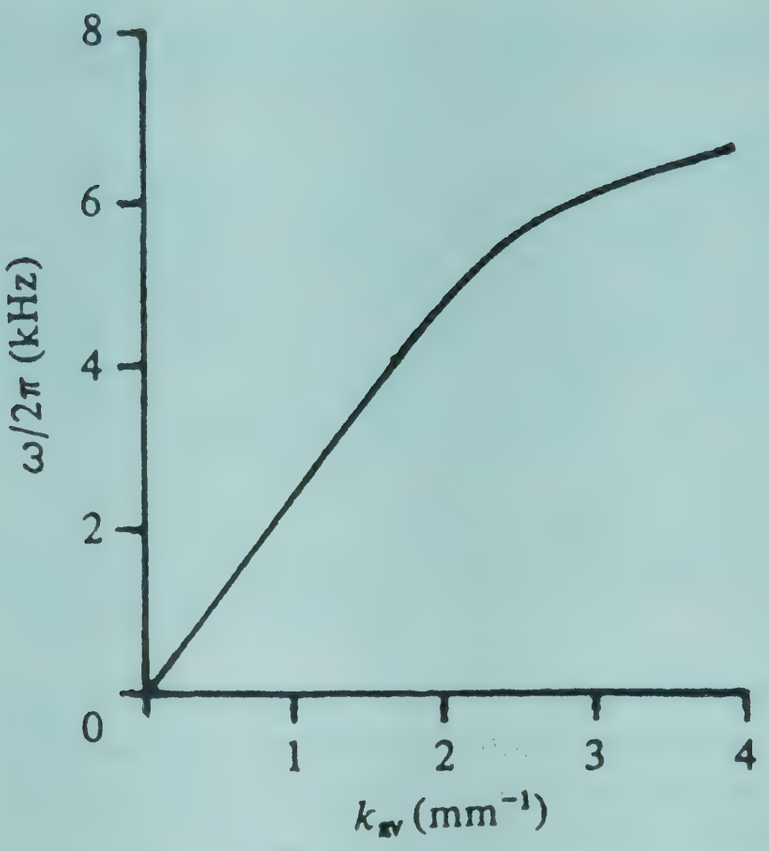


Fig.11. Average wavenumber k_{av} between points 1 and 2, deduced from results in Fig.10 by equation (6), and plotted as abscissa against frequency $\omega/2\pi$ (in kHz) as ordinate. Note that, for a plot of frequency ω against wavenumber k, its slope (7) is the group velocity U; that is, the speed of travel of wave energy.

(that is, as limited by the resolution achievable with measurements 1.5mm apart) is turning over in such a way that its slope,

$$\partial\omega / \partial k = U \qquad \ldots (7)$$

is gradually decreasing to zero.

All of you will recall, of course, the importance of this slope U of such a curve of frequency against wavenumber. It is a velocity usually called the group velocity, also known as the wave-packet speed (in other words, the speed with which a limited group of waves at the frequency in question travels along); and, from the most fundamental point of view, it is the speed of travel of wave energy. Accordingly, we may be very interested in the possibility that this speed of travel U is tending to zero because, when that happens, it may permit the wave energy to "pile up".

Fig.12 shows another way of estimating the velocity of energy propagation U from the data of Rhode (1971). The rate of input of wave energy at the oval window takes the form pressure times volume flow, and this volume flow is proportional to malleus velocity just as is the pressure (see above). So this power input P must be proportional to $(\omega a'_M)^2$, the square of the velocity amplitude of the malleus. On the other hand, the

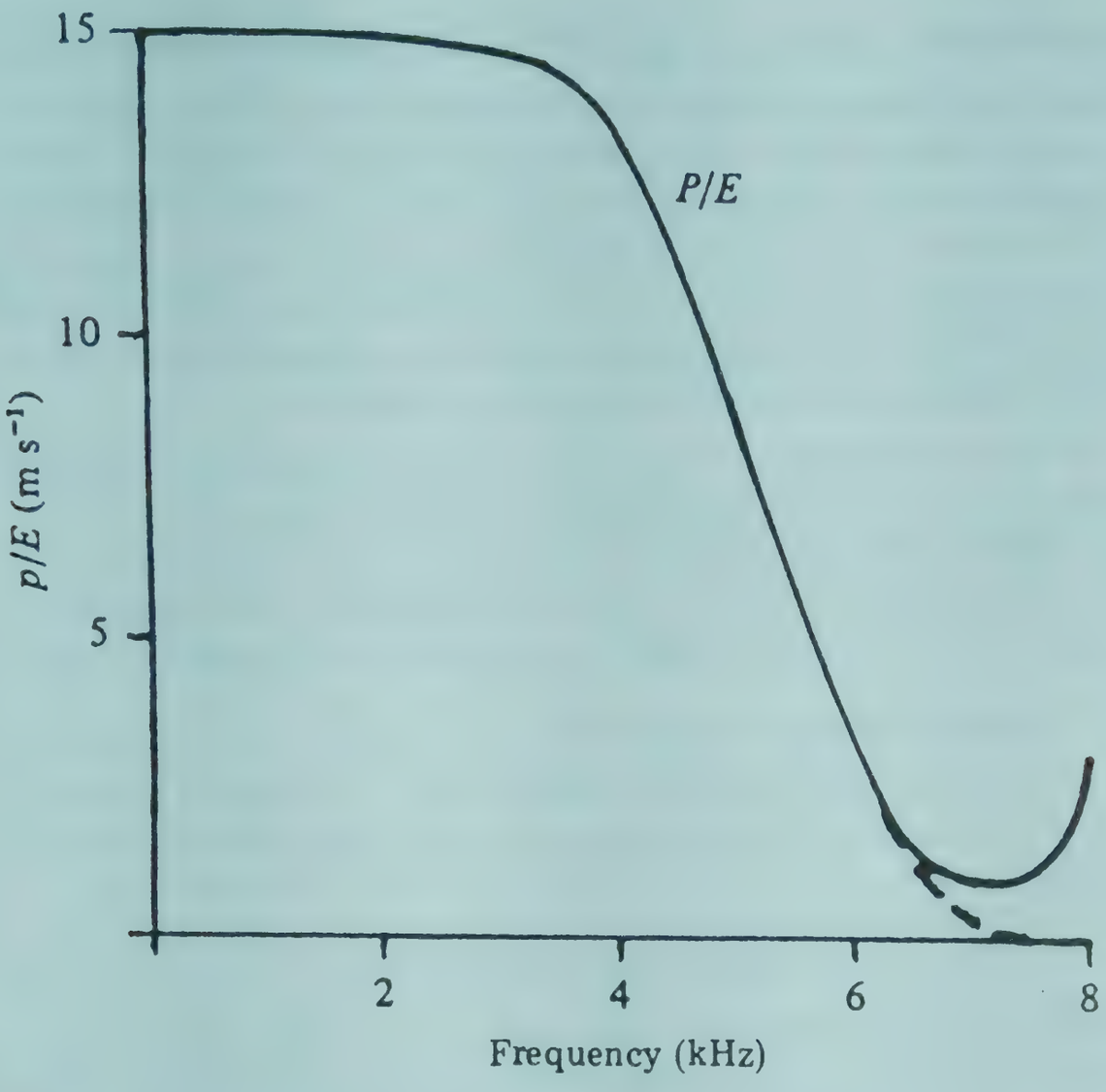


Fig.12. Plain line: ratio P/E of power input P to energy flow E, deduced as explained in the text from the values of a_B/a_M plotted in figure 8. Broken line: possible behaviour of U consistent with this graph of P/E *and also* with equation (8), since dissipation is expected to occur primarily near the resonant frequency $\omega_r(x)$. Note that the graph of U comes in tangentially to zero where $\omega = \omega_r(x)$, with consequences as explained in the text; see especially the consequence (9), dependent on the fact that the ratio of $[\omega_r(x) - \omega]^2$ to $[x_r(\omega) - x]^2$ has the non-zero limiting value $(d\omega_r/dx)^2$.

wave energy per unit length of cochlea must be half kinetic energy and half potential as in any vibrating system; so this energy per unit length E, like the potential energy that represents half of it, must be proportional to the square a_B^2 of the basilar-membrane displacement.

Now, if U is the speed of travel of wave energy, then the rate of flow of energy (that is, the power transmitted by the wave) takes the form UE. This must equal the input power P where no dissipation has occurred, but be less than P whenever some dissipation has occurred. In summary, then,

$$U \overset{\text{no}}{\lessapprox} P/E \text{ where some dissipation has occurred.} \quad \ldots (8)$$

From Rhode's data we can plot P/E, knowing that it is $15ms^{-1}$ at low frequency and varies in proportion (see above) to $(\omega a_M)^2/a_B^2$). Figure 12 gives the resulting graph of P/E (plain line) which is consistent with the idea that the energy propagation velocity U follows a curve (broken line) which comes in tangentially to zero at a certain resonant frequency $\omega_r(x)$depending on the position x of the Mössbauer source. Then U would fall below P/E only in the region where the energy is travelling very slowly and therefore has time to suffer significant dissipation.

This idea (consistent as we shall see with theoretical considerations) that the curve of U against ω comes in tangentially to zero where $\omega = \omega_r(x)$ implies a proportionality of U to $[\omega_r(x) - \omega]^2$. This in turn, if we consider the inverse function $x = x_r(\omega)$ identifying for each frequency ω the place x for which that frequency is the resonant frequency, implies that

$$U \text{ is proportional to } [x_r(\omega - x)]^2 \quad \ldots (9)$$

It follows from Eq. (9) that the "group delay" for waves of frequency ω that is the travel time for wave energy from the base to reach a point x, takes a form,

$$\tau = \int_0^x U^{-1} dx, \text{ which becomes unbounded as } x \to x_r(\omega) \quad \ldots (10)$$

This is fully consistent with the idea that any damping rate however small (for example, energy dissipation by viscous action in thin Stokes boundary layers) has all the time it needs to dissipate the energy[5] .

On the other hand, the very slow speed at which the energy is travelling means that this dissipation all takes place within a very small distance. Actually it is possible as Zweig (1976) showed, using the data of Rhode (1971) with interpolation between data points achieved by very reasonably assuming phase to be a function of ω/ω_r, to reconstruct the travelling wave on a squirrel monkey's basilar membrane excited by a pure tone of 70dB amplitude at 2kHz. Figure 13 shows, both the distribution of displacement (in nanometres) at one instant of time, and the envelope of the travelling wave. It shows the expected very high values of wave velocity and wavelength near the base, together with the gradual slowing down of the wave energy and massive increase of wavenumber as the characteristic place is approached. There the energy "piles up" into a local

---

[5]Here there is a close analogy with the phenomenon of "critical layer absorption" encountered in geophysical fluid dynamics (Booker & Bretherton, 1967; Lighthill, 1978b, 1981).

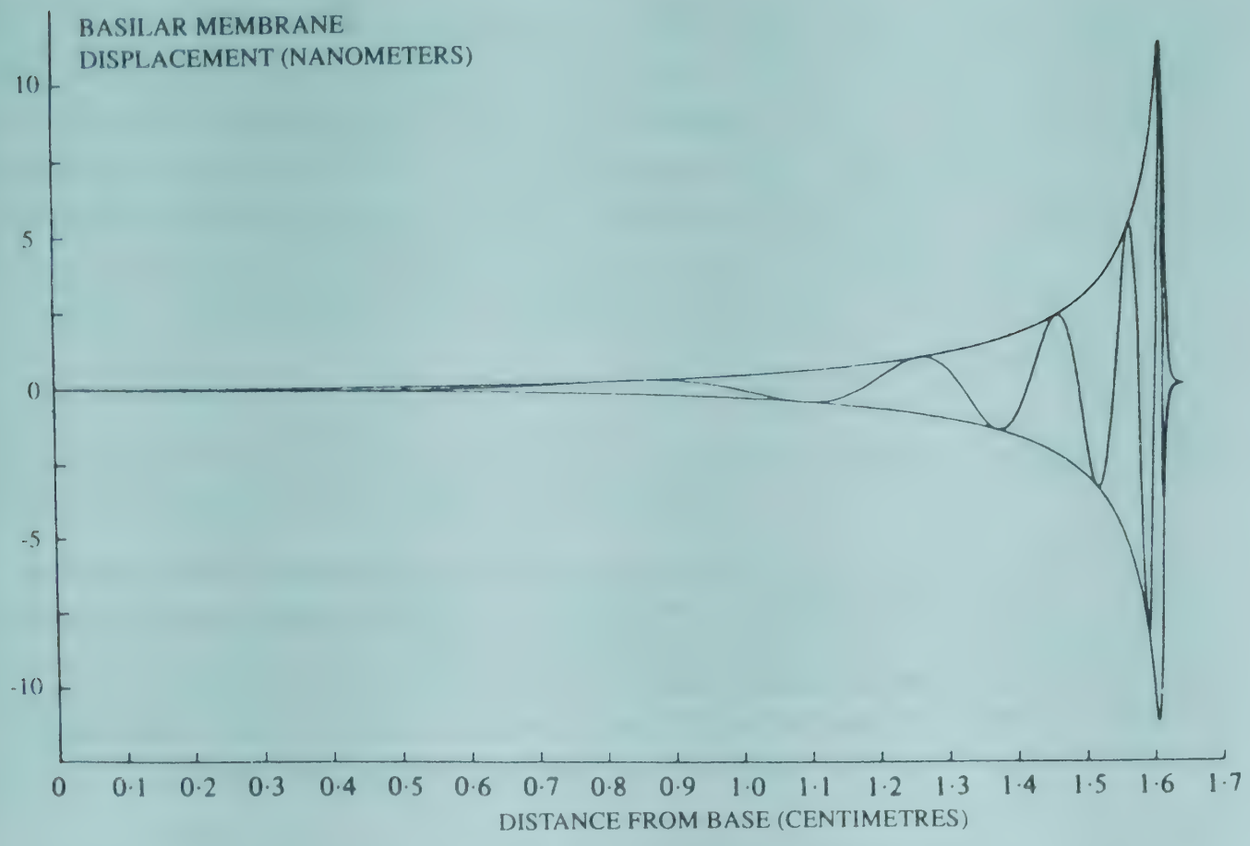


Fig.13. Reconstruction by Zweig (1976) of the travelling waves on a squirrel monkey's basilar membrane excited by a pure tone of 70dB amplitude at a frequency of 2kHz. The displacement curve for the travelling wave is shown at one instant of time (note the transition from very long to very short waves); the envelope of the travelling wave is also given. This reconstruction is computed from measurements of amplitude and phase by Rhode (1971); the assumption used to interpolate between Rhode's measurements at different points of the membrane being that phase is a function of ω/ω_r. Note that the vertical scale is enormously enlarged; being given in nanometres as against a horizontal scale in centimetres.

oscillation at 2kHz before being dissipated in the large time available as the energy propagation speed tends to zero.

5. Macromechanical theory

Up to this point I relied entirely on good experimental data to establish (section 4) the presence of wave dispersion and the character of such dispersion for waves propagated in the macromechanical system which comprises the basilar membrane and the fluid in the two adjacent tubes which it separates. Any acceptable macromechanical theory should account for this dispersive behaviour.

In particular, it should be compatible with the suggestion that the dispersion curves of frequency against wavenumber, at different positions on the basilar membrane represented by their different distances x from the base, are expected to take forms as shown in Fig.14 (Lighthill, 1983); with straight line portions (representing nondispersive behaviour) at low

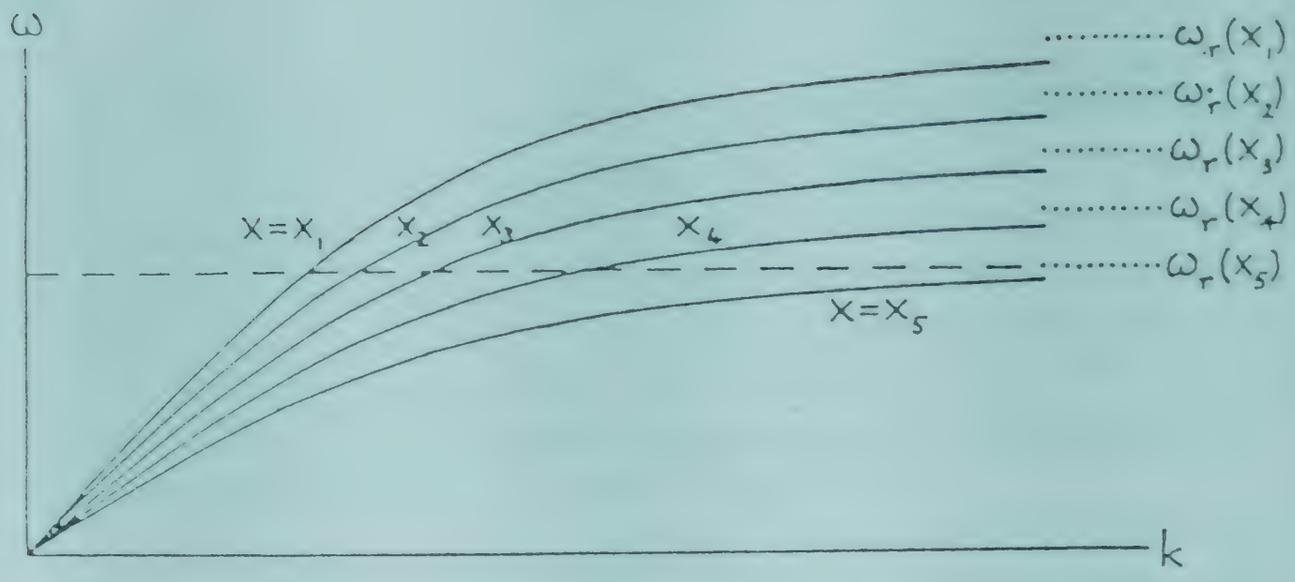


Fig.14. The data of Rhode (1971) suggest that dispersion curves at different places x_1, x_2, x_3, x_4, x_5 on the basilar membrane should take forms like those shown here, asymptotic to the frequencies $\omega_r(x_1)$, $\omega_r(x_2)$, $\omega_r(x_3)$, $\omega_r(x_4)$, $\omega_r(x_5)$, for which they are the characteristic places. (Here, each curve is a natural extension of Fig.11 as the resonant frequency.)*Broken line*: For a signal of given frequency ω, this traces how k increases to very large values (while the slope $\partial \omega/\partial k = U$ tends to zero) at the characteristic place for that frequency].

wavenumber, followed by reductions in slope as wavenumber increases until the slope finally tends to zero at an asymptote corresponding to the place in question. At each position x, this shape of curve represents an extension of the experimental curve in Fig.11 to include the limiting case of large wavenumber, where the energy propagation velocity (that is the group velocity, given by the slope of the curve) tends to zero. Conversely, the broken line indicates how, for a vibration of fixed frequency ω travelling along the cochlea, the same reduction in energy propagation velocity to zero occurs as x increases to the value $x_r(\omega)$ for which $\omega_r(x) = \omega$; in other words as the acoustic signal propagates along the cochlea as far as the characteristic place.

Specifically, the asymptotic behaviour in Fig.14 needs to take the form

$$\omega_r(x) - \omega \text{ proportional to } k^{-1} \qquad \ldots (11)$$

in order that

$$U = \partial\omega/\partial k \text{ is proportional to } k^{-2} \text{ and so to } [\omega_r(x) - \omega]^2 \qquad \ldots (12)$$

as suggested by Fig.12. This in turn will lead to the conclusions (9) and (10), allowing energy at frequency ω to "pile up" at the characteristic place $x_r(\omega)$.

These essential demands (see Fig.14 and equation (11)) from a macro-mechanical theory underline the importance of two properties of the

basilar membrane and of its adjoining fluid media which I highlighted in my initial outline (section 3). First, there is the crucial experimental discovery by Voldrich (1978) that the mechanical behaviour of the basilar membrane *in vivo* is highly anisotropic (see also Voldrich, 1983), to the extent that neighbouring short sections of the membrane vibrate almost independently of one another. This means that, at each place x on the membrane, we can independently define a stiffness

$$s(x), \qquad \ldots (13)$$

in terms of an overall pressure excess in the scala tympani divided by the resulting increase in its cross-sectional area.

Here, in the language of the general theory of vibrations, we are specifying the relevant amount of *displacement* in the mode in which the basilar membrane tends to bend under pressure by means of a "generalised coordinate" h defined as the resulting increase in cross-sectional area. The required pressure difference is then

$$s(x)h, \qquad \ldots (14).$$

so that the associated potential energy per unit length of membrane takes the form expected,

$$\tfrac{1}{2}\, s(x)h^2, \qquad \ldots (15)$$

in terms of the appropriate generalised stiffness[6].

We can also expect the corresponding kinetic energy per unit length to take a similar form

$$\tfrac{1}{2}\, (m_b + m_f)\dot{h}^2, \qquad \ldots (16)$$

in terms of a generalised velocity $\dot{h}$ (the time derivative of the generalised coordinate). In expression (16) the part

$$\tfrac{1}{2}\, m_b(x)\dot{h}^2, \qquad \ldots (17)$$

represents the kinetic energy of the basilar membrane itself together with those solid components of the organ of Corti that vibrate with it; while m_f is the added inertia of the corresponding movements induced in the fluid.

And this brings me to the second point which I highlighted in my initial outline (section 3): the fact that, as waves in the membrane become shorter, the fluid motions become more and more concentrated near it,

[6]Strictly, equation (15) is the true definition of stiffness; being compatible with the informal definition (13) provided that "overall" pressure excess is interpreted as a weighted mean of the distribution of pressure excess associated with the membrane's motions (the weighting factor being membrane displacement).

within a distance from it which I referred to there as a wavelength divided by 2π but which it is more correct to write in terms of the wavenumber k as

$$k^{-1} . \qquad \ldots (18)$$

There is a fundamental reason for this concentration, of course: outside very thin Stokes boundary layers (of thickness just a few micrometres at these frequencies), the fluid motions in any cross-section satisfy Laplace's equation; which, for waves propagating in the x-direction with wavenumber k, takes the form

$$\frac{\partial^2\phi}{\partial y^2} + \frac{\partial^2\phi}{\partial z^2} - k^2\phi = 0 \quad . \qquad \ldots (19)$$

Disturbances satisfying equation (19) fall exponentially like e^{-kr} with distance r from any source (such as the membrane's movements); so that their kinetic energy penetrates a distance k^{-1} into the fluid as in (18). It follows that the added inertia m_f of fluid in expression (16) takes a form

$$m_f(x, k) \qquad \ldots (20)$$

dependent, not only on the position x, but also on the wavenumber k; while for large k, furthermore, it follows that

$$m_f \text{ is proportional to } k^{-1} \qquad \ldots (21)$$

Now the general theory of vibrations permits us, for a vibration where equations (15) and (16) specify the potential and kinetic energy, to write down the square of the radian frequency ω as the ratio of stiffness to inertia giving

$$\omega^2 = \frac{s(x)}{m_b(x) + m_f(x,k)} \qquad \ldots (22)$$

(Essentially, this condition ensures that the average values of (15) and (16) are equal.) For extremely large k, we have seen that m_f tends to zero so that

$$\omega \text{ asymptotes to } \omega_r(x) = [s(x)/m_b(x)]^{1/2}, \qquad \ldots (23)$$

a resonant frequency depending on the local properties of the membrane alone[7]. For moderately large k, on the other hand, the correction $m_f(x,k)$

[7] Here, the variation of $\omega_r(x)$ along the cochlea by two orders of magnitude (section 2) is of course primarily related to the variation of the stiffness s(x) by close on four orders of magnitude.

in (22) is proportional to k^{-1} (see (21) ,giving the desired prediction (11) that $\omega_r(x) - \omega$ is also proportional to k^{-1} .

It is important that this satisfactory conclusion has been arrived at without needing to know an exact bending mode for the basilar membrane (in the form of the distribution of displacement across its width). Whatever that mode may be, the fluid inertia takes the asymptotic form (21) so that the dispersion curve has the asymptotic property (11) which gives the cochlea its frequency selectivity, with energy in each frequency "piling up" at the characteristic place for that frequency.

For completeness, I show in addition that equation (22) is compatible with the requirement that the wave becomes nondispersive for small k, when the wave fills each tube and also becomes approximately a *longitudinal* wave. If the longitudinal volume flow in one tube (say, the scala tympani) has amplitude J then the equation of continuity tells us that for wavenumber k the amplitude $\dot{h}$ of rate of change of area must be kJ. Accordingly, the total kinetic energy per unit length $[(\frac{1}{2}\ \rho A^{-1}J^2]$ in the scala tympani, of area A, plus an equal amount in the other tube is

$$2(\tfrac{1}{2}\ \rho A^{-1}J^2] = \frac{1}{2}(2\rho A^{-1}k^{-2})\dot{h}^2 \qquad \ldots (24)$$

This gives

$$m_f \sim 2\rho A^{-1}k^{-2} \text{ for small k,} \qquad \ldots (25)$$

allowing the corresponding asymptotic form of equation (22) to be

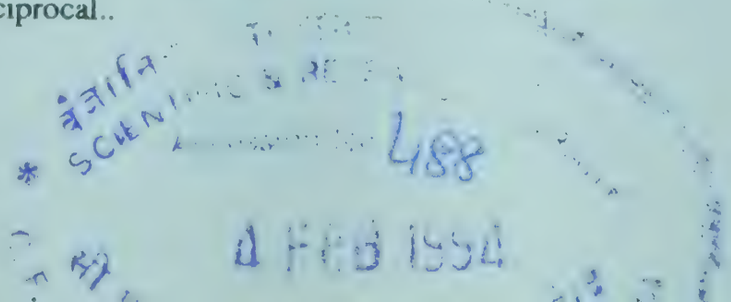
$$\omega \sim [As(x)/2\rho]^{1/2}k, \qquad \ldots (26)$$

exactly as in the nondispersive theory[8].

I have demonstrated quite briefly, then, how some simple principles from the general theory of vibrations and waves, when applied to cochlear macromechanics, yield theoretical conclusions that are fully compatible, not only with our overall knowledge of frequency selectivity in the cochlea, but also with good experimental data on basilar-membrane response at moderate sound levels like the observations at 70dB made by Rhode (1971). And we have seen that, qualitatively, such theoretical conclusions can be drawn without our needing to have precise knowledge of mode shapes at the different cochlear cross-sections. Now, rather than

---

[8]Indeed the "equivalent distensibility" as defined in a footnote (no.3) in section 3 is readily seen to have $\frac{1}{2}$ As(x) as its reciprocal..



trying to go into the sort of further detail that necessarily depends on making assumptions about those mode shapes, and instead of explaining how the contents of this section are related to the detailed methodology pursued in various studies which introduce such assumptions [studies that were brilliantly pioneered in the San Francisco area by Charles Steele (1974)]. I shall consign all of that to an extended footnote in my published text[9] in order to be able to move forward at once to the challenges posed by more recent experimental data.

---

[9]All of the theoretical studies which make a serious attempt to model the distribution of disturbances over the width as well as the depth of the cochlear cross-section make use of exactly the same approximation. Some, with good historical justification, name it the Liouville-Green approximation (Viergever & Diependaal, 1983) and some, with advantages of brevity, the WKB approximation (Steele, 1974; 1976; Steele & Taber, 1979). Some call it high-frequency asymptotics (Holmes, 1982) and some energy-flow theory [Lighthill (1981, 1983), who gives a detailed account of the method outlined in section 5]. But all these asymptotic theories in the high-frequency limit are mathematically equivalent, as shown for example in my book (Lighthill, 1978b). Certain complications in the earlier papers which modelled the basilar membrane as an isotropic elastic plate were made unnecessary after the work of Voldrich (1978) which showed how neighbouring sections of membrane move almost independently, and models incorporating this discovery exhibited sharper tuning. An excellent investigation of the validity of the high-frequency approximation was offered by Steele & Taber (1979) who compared an exact calculation, made by Allen (1977) of a model involving purely two-dimensional motions (uniform across the width of the cochlea), with their own WKB analysis of the same model and found practically identical results.

As the characteristic place is approached, the waves get shorter and shorter, but a critic might ask if the approximation nevertheless gets worse because the gradient of wavenumber is so steep. Actually, in the evaluation of the approximation, the criterion for adequate accuracy is the relative change in the wavenumber k in a distance given by the inverse wavenumber (18); and, as Lighthill (1981) points out, this quantity $k^{-2}\,dk/dx$ does not increase (it is asymptotically constant) as the characteristic place is approached.

Strictly, the characteristic place for frequency ω, if defined as the place where the amplitude reaches its maximum, appears just ahead of the position $x_r(\omega)$ where the resonant frequency is equal to ω. The theories indicate, in fact, that damping has dissipated all the energy in the wave before the place $x_r(\omega)$ is reached; albeit (as suggested by Fig.13) "only just" before.

At rather low frequencies (less than 1kHz) the high-frequency approximation does not retain its validity all the way to the base: Lighthill (1981) shows that it ceases to be valid for $k < 0.2\text{mm}^{-1}$. In this region, however, the waves are strictly nondispersive and longitudinal so that a simple match of that approximate solution to the exact one-dimensional stiffness-dominated solution of Zwislocki (1948) produces results uniformly valid throughout the cochlea.

PART 3. COCHLEAR MICROMECHANICS

6. Nonlinear response at low amplitudes

The final part of my lecture, concerned with cochlear micromechanics, begins then with sections describing two quite separate groups of relatively newer experiments on the response of the cochlea as a whole which have each helped, on analysis, to bring about an improved recognition of the role that its micromechanics must be playing. After them I shall briefly indicate how brilliant sub-microscopic studies have recently gone far towards uncovering detailed mechanisms at the cellular level which seem to be mediating that response.

The first of those groups of experiments deals with the most intriguing aspects of *nonlinear* response by the cochlea to acoustic signals of various amplitudes. And perhaps I should explain at once that in using the phrase "most intriguing aspects", I am completely excluding all consideration of the cochlea's nonlinear response at high amplitudes. After all, mechanics teaches us *that* essentially every vibrating system exhibits nonlinear response to forcing at high amplitudes. The cochlea is no exception; but details of *that* response in the cochlea, being in no way extraordinary, may appropriately be categorised as "less intriguing" - besides being inappropriate for inclusion in an account of hearing sensitivity.

By contrast, we all have to sit up and take notice when good experimental data indicate that some particular feature of the response of any system becomes more and more nonlinear as the amplitude gets less and less! Of course early data making such an improbable suggestion may quite reasonably be queried as of uncertain validity[10], but when a whole sequence of data of demonstrably increasing accuracy has forced us to so unusual a conclusion, we have to start taking it very seriously indeed.

In the cochlea, the feature of the basilar membrane's response which becomes more and more nonlinear at low amplitudes (in other words, at low sound levels) is the height of the peak of the response curve - that is the curve of amplitude ratios, as plotted in Fig.8 for a particular squirrel monkey subjected in the experiments of Rhode (1971) to pure tones at a

[10]This happened, for example, to some data along these lines obtained in squirrel monkeys (Rhode & Robles, 1974) with, however, sound levels not reduced below 70dB. Later, on the other hand, the very precise and comprehensive measurements in chinchilla by Robles, Ruggero & Rich (1986) produced fully convincing conclusions, along closely the same lines as those of Sellick, Patuzzi & Johnstone (1982) in guineapig which are described below.

sound level of 70dB. It is not any other feature of the curve that responds nonlinearly; it is just the height which increases more and more as the sound level is reduced. And, once again, anyone familiar with how extraordinarily various systems have been refined during two hundred million years of mammalian evolution by natural selection has to sit up very respectfully indeed in front of well attested evidence of anything so potentially valuable to an animal's survival as an auditory response curve which, around its peak, becomes more and more sensitive as sound levels get lower and lower. Immediately one is tempted to puzzle over what special sort of mechanism might have evolved to allow such an advantageous increase of sensitivity at low stimulus levels.

This was the point reached within the specialist world of auditory biophysicists in 1982, as a result of two papers out of which I shall describe the one (Sellick, Patuzzi & Johnstone, 1982) which used the Mössbauer-source technique that I have already described to you, but used it this time in guineapig and with several important refinements. A quite different experimental approach (based on laser interferometry) which I've no time to outline was used by Khanna & Leonard (1982) to arrive, independently, at essentially the same conclusions in cats.

Sellick, Patuzzi & Johnstone (1982) at the University of Western Australia had, in particular, perfected the surgical procedures for Mössbauer source implantation on the basilar membranes of their guineapigs so as to minimise cochlear damage as assessed in various ways. Then in a very extensive series of experiments, they found that peak sensitivities were enormously enhanced (in the nonlinear manner which I have described) when two conditions were satisfied:

(a) cochlear damage minimised;

(b) experiment performed immediately on completion of surgery.

Like many authors they show their response curves in a form which is like Fig.8 but upside down; they plot the sound pressure level required to produce a given vibration of the basilar membrane (rather than the other way round; but, on a logarithmic scale, this simply inverts the curve).

The curve so obtained under conditions (a) and (b) was shown already in Fig.7. The sharpness of the inverted peak is here most striking, both in its own right and also because it is just as sharp as a typical tuning curve for individual auditory nerve fibres of guineapigs as obtained in other experiments. This latter curve plots as a function of frequency the acoustic threshold for the nerve fibres; in other words the minimum sound level needed to produce a signal in the nerve fibre. The similarity between the

two curves in Fig.7 is very marked and is consistent with the idea that basilar-membrane motions are directly responsible for stimulating neural activity (through, of course, the transduction mechanism in inner hair cells).

One might perhaps be tempted to infer from these data that the macromechanical process in the cochlea which, as I showed earlier, gives basilar-membrane vibrations their excellent frequency selectivity, is also responsible for the healthy ear's very low hearing thresholds. That view is untenable, however, because such a passive mechanical process is purely linear at low amplitudes, producing admittedly a sharp resonance peak, but one whose height is limited by a level of energy dissipation which cannot fall below that associated with the viscosity of the fluid.

Furthermore, Sellick, Patuzzi & Johnstone (1982) produced clear evidence that something more than a passive mechanical response is present. Essentially, they demonstrated what they called the "physiological vulnerability" of any sharp tuning curve like that of Fig.7. When either of conditions (a) or (b) is not satisfied, there is a loss of sharpness amounting to as much as 20, 40 or even 60dB. They found too the same loss of sharpness in the neural threshold, as approximated by the threshold sound level that was observed to produce a measurable action potential within the cochlea. This is consistent with the idea that it is basilar-membrane movement which stimulates the generation of neural activity by inner hair cells.

They emphasized, however, how hard it is to account for the observed physiological vulnerability (and, especially, the loss of sensitivity at quite small elapsed times after completion of surgery) by means of any change in passive mechanical properties. They felt constrained rather to advance what we shall see was a most important suggestion; namely, that an active process which requires metabolic energy amplifies "in some unknown way" the signal at the resonant frequency.

Finally, before I go on to outline certain quite different evidence pointing to the same conclusion, I will show yet another very interesting discovery by Sellick, Patuzzi & Johnstone (1982). In all of their experiments, including some that satisfied conditions (a) and (b) rather closely and others departing from one or both of them by varying amounts, they tabulated the threshold sound levels for which a cochlear action potential was observed and they found that in every case it corresponded closely (see Fig.15) with the sound level producing a basilar-membrane velocity of 0.04mm/s. This is consistent with the idea that an inner hair

cell is stimulated when its stereocilia respond to a certain magnitude of velocity distribution in the fluid and in associated components of the organ of Corti.

Once again, however, I emphasize the other striking feature of these experiments: both of the two thresholds that are plotted against one another in Fig.15 (that for producing a cochlear action potential and that for generating a basilar-membrane velocity of 0.04mm/s) are changing by up to 60dB in data obtained *either* when conditions (a) and (b) are satisfied

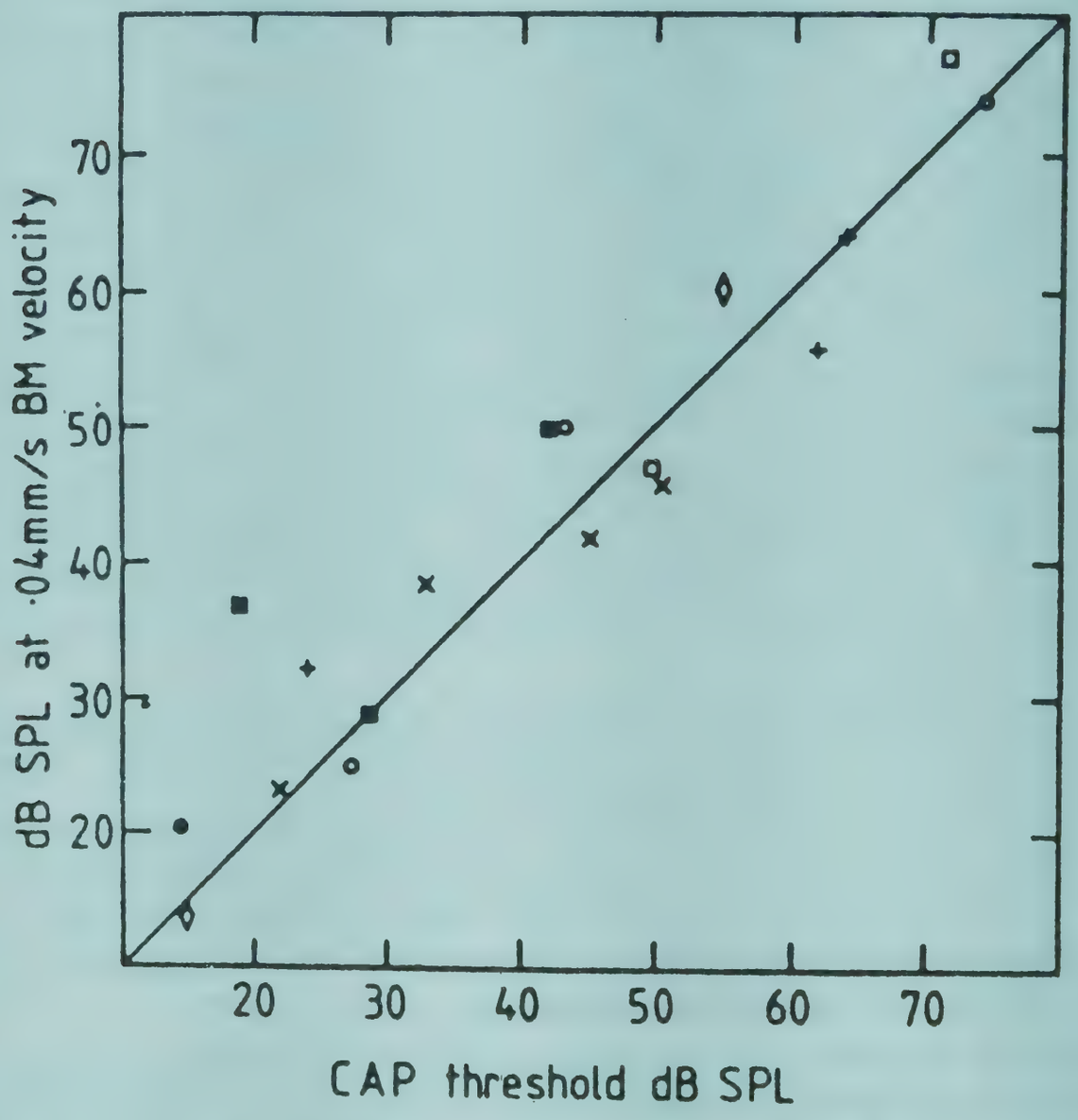


Fig.15. For a range of different experiments by Sellick, Patuzzi & Johnstone (1982), where *either* conditions (a) and (b) were satisfied or one *or* both conditions failed to be satisfied to varying degrees, two different threshold sound pressure levels (at resonant frequency) are here plotted against each other: (i) the threshold for producing a basilar membrane (BM) velocity amplitude of 0.04mm/s and (ii) the threshold for exciting neural activity as indicated by the production in the cochlea of a measurable action potential (CAP). Essentially, both thresholds rise in parallel as cochlear function is degraded.

Figures 7 and 15 are consistent with a view that the transduction mechanism in inner hair cells rather directly converts basilar- membrane vibration velocities into neural activity.

or after various degrees of departure from one or both of them. This feature suggests that the nonlinear properties of the cochlea, which allow a progressive sharpening of its response curve at low sound levels, are mediated by a physiologically active process that is vulnerable to surgical insult and trauma.

7. Evoked otoacoustic emissions

In the meantime the remarkable studies of David Kemp in London had demonstrated (Kemp, 1978, 1980) by noninvasive experiments in both human subjects and experimental animals that pure-tone signals are able to evoke from the ear an acoustic response, which takes the form of an emission of sound from it at the same frequency, and which is both nonlinear and physiologically vulnerable. This response, known as the otoacoustic emission, provides independent evidence that an acoustic signal stimulates in the cochlea an active process, utilising metabolic energy and involving mechanical vibrations at the stimulus frequency.

In Kemp's experiments as well as in other scientists' more recent observations of the same phenomenon a microphone, placed in the outer ear canal, measures the emission from the ear (Fig.17) which emerges, with a short time lag, after the ear is stimulated by a pulse of pure-tone sound (shaped carefully to minimise energy in sideband frequencies). The clearest evidence that the emission comes from a physiologically active process in the cochlea is provided by the fact that emission is completely suppressed by the action of drugs known to inhibit cochlear activity.

Kemp (1980) showed, furthermore, that the emission observed is consistent with a travelling-wave interpretation as follows. This is that the stimulus given to the eardrum excites, through the linkage mechanism in the middle ear (Fig.1), the usual travelling-wave in the cochlea (Fig.13) which in turn, when it reaches the characteristic place, stimulates the emission of a backward travelling wave. (Needless to say, the macro-mechanical system of travelling dynamic waves described in section 5 admits their propagation from base towards apex or in reverse[11]). That backward travelling wave exerts a pressure on the oval window which,

[11]Occasional suggestions to the effect that this is a system favouring just the one-way transmission of waves are, of course, quite unfounded. For example, the experiments reported by Wever & Lawrence (1954) on a cat's cochlea given acoustic stimulation at the apex produced results entirely consistent with the view (Lighthill, 1981) that this stimulation generated a "fast wave" (see beginning of section 3) travelling from apex to base; where, because of the differing impedances of the round and oval windows, it generated an apically travelling "slow wave" such as they observed.

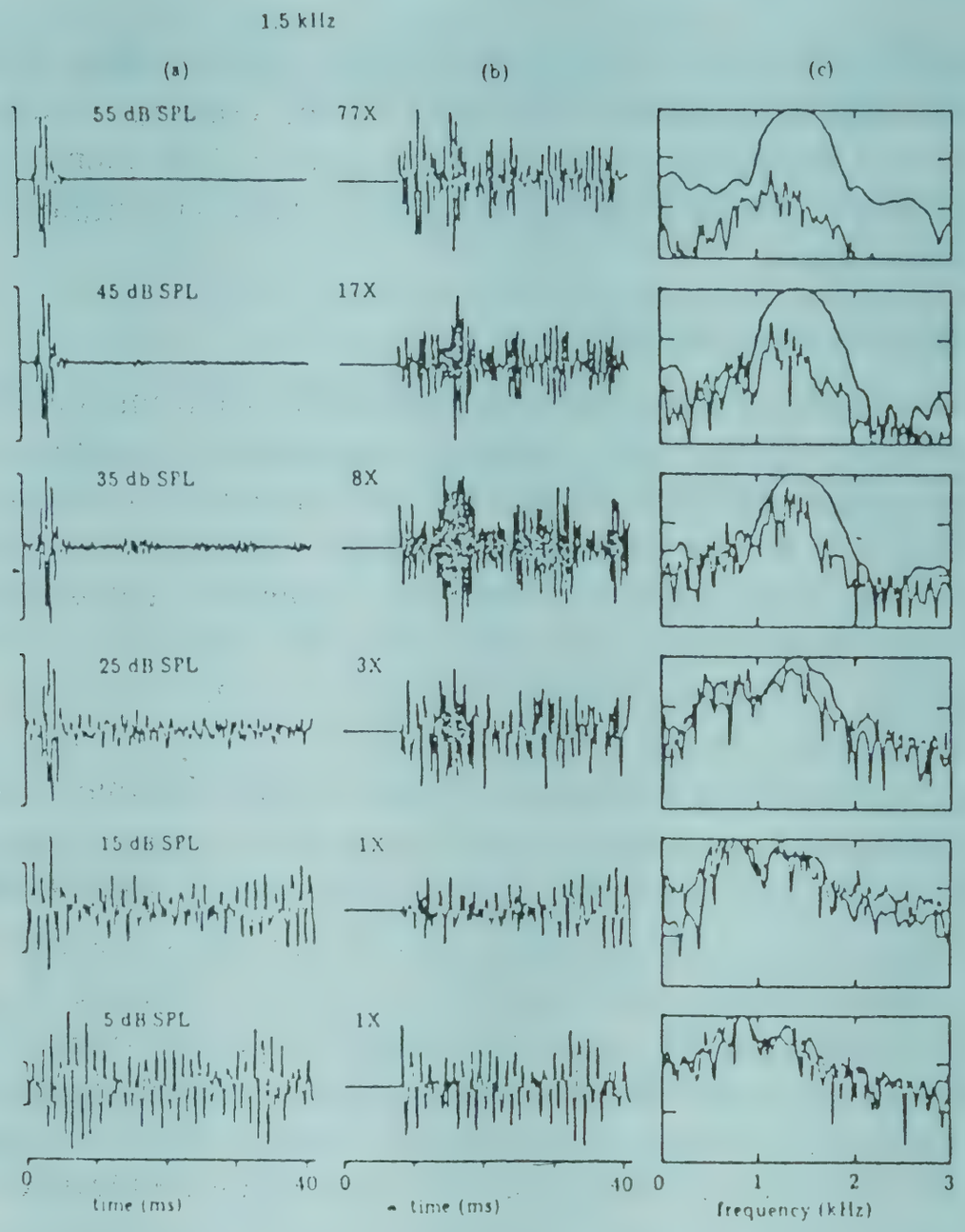


Fig.16. Data by Norton & Neely (1987) on otoacoustic emissions evoked by pure-tone bursts at six different sound pressure levels (from 55dB to 5dB).

Column (a): the unprocessed wave form (combining the pure-tone burst and the emission) as recorded by the microphone in the outer ear canal. This column demonstrates the extreme nonlinearity of the phenomenon. Note that at 55dB only the stimulating pure-tone burst is visible on the record (but see Column (b) for a scaled-up version). As the sound level is reduced, all six records show the stimulating pure-tone burst on the same scale, but the evoked otoacoustic emission becomes more and more evident, and at 5dB has an amplitude as high as that of the stimulus (and, indeed, a bigger energy because of the emission's extended duration).

Column (b): the same waveforms after the stimulus has been subtracted out and the remainder, representing the otoacoustic emission, scaled up by the factors shown (77 x means a linear scaling by a factor of 77 times).

Column (c): spectra for the stimulating pure-tone burst (smooth line) and for the evoked otoacoustic emission. At the lower sound levels they practically coincide. Note that the time delay for emission is less for higher-frequency components of the signal (with a characteristic place nearer the base) than for lower- frequency components. It is this variability in time delay which spreads out the duration of the emitted signal.

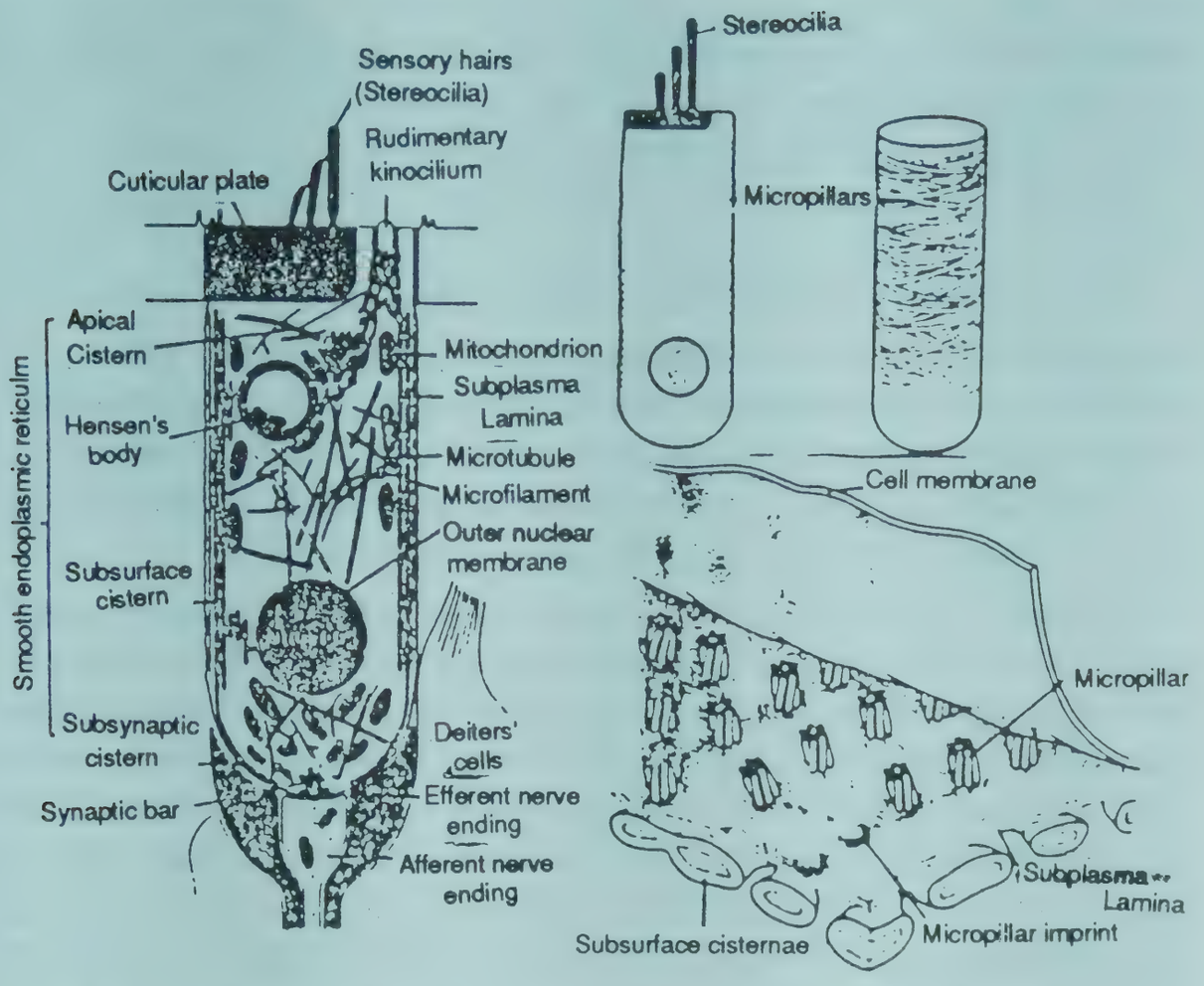


Fig.17. A schematic representation of the sub-microscopic structure of the outer hair cell (Lim, Hanamura & Ohashi, 1989; reproduced here by courtesy of Plenum Press).

Left: overview of the substructures of the outer hair cell.

Top right: arrangement of "micropillars" in a "criss-cross" assemblage of parallel arrays.

Bottom right: artistic conception of the subplasma lamina with micropillars which bridge between the outer-hair-cell plasmalemmas and the subsurface *cisternae*.

through the same linkage acting in reverse, causes the eardrum to generate the signal observed in the outer ear canal.

Any engineer, of course, might be tempted to diagnose such a response as a simple reflection such as might readily be supposed possible in a ducted system. But this is an untenable interpretation: not only would any passive reflection process persist after the drugs I referred to had acted, but also it would (at low amplitudes) be a linear process, with a reflection coefficient independent of amplitude. Yet Kemp (1980) estimated the effective reflection coefficient as very much bigger at low sound levels (it is about 1 near the threshold of hearing, for example) than it is at higher levels.

Other research workers (Wilson, 1980; Wit & Ritsma, 1980, 1983) have confirmed the marked nonlinearity of the phenomenon and its

dependence on a physiologically active cochlear process. In addition, some recent studies by Norton & Neely (1987) and by Neely, Norton, Gorga & Jestendt (1988) in human subjects have allowed rather precise comparisons between the measured time lag in the otoacoustic emission produced by a pure-tone pulse and their determinations of the "latency" or time delay (more strictly, the group delay) for the travelling wave generated by such a pure-tone pulse to reach its characteristic place. Their results are consistent with the view that the otoacoustic emission is stimulated when this wave reaches its characteristic place and that the backward travelling wave then takes the same time to reach the base and excite the eardrum; so that the emission time lag is twice the latency.

I should emphasize that otoacoustic emissions are important for many reasons, including (in human patients) various diagnostic purposes. In this lecture, however, I am concentrating on their relevance to the interpretation of that nonlinear response of the cochlea at low sound levels which we have seen (section 6) to be so important for hearing sensitivity[12]. This relevance depends on an essentially unavoidable inference that the process which acts as a source of otoacoustic emissions at the stimulus frequency must take the form of a vibration at that frequency. Evidently, the fact that a low-level acoustic signal stimulates a physiologically active process involving such a vibration may be extremely important in enhancing, at low sound levels, the peak amplitude of basilar-membrane vibrations. In the final section of this lecture I shall outline evidence supporting the idea that a particular active mechanism, involving vibrations of the outer hair cell, is brought into play at low stimulus levels to amplify those vibrations of the basilar membrane which in turn, through the transduction capability of inner hair cells, generate signals in auditory nerve fibres.

Last year, many specialists on cochlear mechanisms from all over the world gathered together in England at the University of Keele for prolonged and stimulating evaluation of this idea (later published under the able editorship of Wilson & Kemp 1989). A special guest of honour at this meeting was that very wide- ranging scientist Thomas Gold who in 1948 won a Fellowship of Trinity College Cambridge with a Dissertation arguing that the cochlea must incorporate such a mechanical feedback system (see also Gold & Pumphrey, 1948; Gold, 1948). In the late 1940s, regrettably enough, hearing specialists had shown themselves unready to

[12]This is the reason for referring only to evoked emissions, whereas the(rather more sparsely observed) spontaneous emissions, though important in other contexts, are excluded from the present discussion.

take seriously Gold's revolutionary views, and the resulting discouragement led to his moving into other fields such as astrophysics where his creative influence has, of course, been profound. We were happy at Keele to make some retrospective amends to Thomas Gold by displaying an impressive body of evidence to the effect that he had been right all along.

8. Micromechanical interpretations of hearing sensitivity

Now, very briefly indeed, I shall speak about the mammalian cochlea at the microscopic and sub-microscopic levels of resolution. Its hair cells have, as mentioned many times, become highly differentiated (Lim, Hanamura & Ohashi, 1989), with the inner hair cells specialising in the transduction function whereby a mechanical vibration produces neural activity in afferent nerve fibres. Before leaving them I will remark that their positioning between the basilar membrane and the overarching tectorial membrane (Fig.4) makes their stereocilia sensitive *both* to fluid motions, which tend to be concentrated (section 5) in this area of the scala media, and to *relative* movements of the tectorial and basilar membranes. Recent evidence on elastic properties of the tectorial membrane *in vivo* (Zwislocki, 1988) supports the earlier suggestion of Zwislocki (1986) to the effect that the vibration of the basilar membrane may, by inducing associated movements in the tectorial membrane, magnify the resulting effect on the stereocilia of inner hair cells, and so on the consequent neural activity. I emphasize, on the other hand, that absolutely no evidence for any active motility has been found in inner hair cells (Forge, 1989).

By contrast, a remarkable high-frequency motility has been demonstrated to be present, at acoustic frequencies, in meticulous experiments made on isolated outer hair cells from a guineapig cochlea by Ashmore (1987, 1989) at Bristol; who measures the precise amplitude of fluctuating elongation and foreshortening of the cell, and the resulting forces exerted by it, in response to specific fluctuations of transmembrane potential. Zjenner, Reuter, Plinkert, Zimmerman & Gitter (1989) obtained similar results in isolated cells, and were also able to make experiments in an intact organ of Corti cultured under conditions that allowed both inner and outer hair cells to produce physiologically normal electric potentials. They found, in this preparation, motile responses of the three rows of outer hair cells at frequencies up to 15 kHz.

Sustained motility at such a high frequency seems to be a rather unique property of outer hair cells which has not been observed in other types of mammalian cell and which poses, therefore, some challenging problems of explanation to cell biologists. I must emphasize that normal motility

mechanisms, such as those found in our muscles that are based on relative movements of actin and myosin fibres, powered by the dephosphorylation of adenosine triphosphate, can be excluded completely as unable to be effective at audiofrequencies.

Against this background various subtle theories have been advanced to account for the very specialised high-frequency motility found in outer hair cells. While I am far from qualified to attempt to describe any of these, I will remark that all are founded on exceptional features of the anatomy of outer hair cells at the ultrastructural level (Lim, Hanamura & Ohashi, 1989), including special characteristics of the plasma membrane and the presence beneath it of several layers of subsurface *cisternae* (Fig.17), anchored to the membrane by "micropillars".

Most auditory physiologists now believe, then, that those active elements in the cochlea which vibrate so as to generate otoacoustic emissions are the outer hair cells. They also believe that in healthy ears these outer-hair-cell vibrations are a source of positive mechanical feedback, helping to amplify the basilar-membrane vibrations at low sound levels. In this context, the function of the efferent nerve fibres carrying signals to the outer hair cells is interpreted as a mechanism for "setting" the feedback control system...

All of these views are consistent with observations to the effect that a variety of drugs which are known to produce substantial loss of hearing acuity are also known to act directly upon the outer hair cells. Perhaps I might add here, in the presence of many engineers and acousticians who are concerned about deleterious effects of very high noise levels on hearing acuity, a remark that it is outer-hair-cell function which is thought to be adversely affected as a result of noise exposure at such very high levels.

Finally, I emphasize that this lecture on the biomechanics of hearing sensitivity has been concerned, not with how the brain in man and other mammals analyses the data coming to it along auditory nerve fibres, but with the initial capture of that data in the cochlea. The brain, needless to say, can produce all its miracles of interpretation only where it works on good initial data. For frequency selectivity these depend on some remarkable properties of the cochlea as a passive macromechanical system, comprising the basilar membrane with its steeply graded stiffness distribution vibrating within the cochlear fluids. But the biomechanics of hearing sensitivity to low levels of sound (at any particular frequency) calls also nto play an active micromechanical system, which during the past few

years has progressively been identified as located in the outer hair cells; and which, through a process of positive feedback, amplifies (in healthy ears) that basilar membrane vibration. This in turn offers the inner hair cells an enhanced signal at low sound levels, so that the threshold at which they can generate activity in auditory nerve fibres is, in consequence, very substantially lowered. I believe that my distinguished audience may have been excited and heartened by both aspects of our improved appreciation of the biomechanics of hearing sensitivity.

ACKNOWLEDGEMENT

I am most warmly grateful to my colleague Professor David Kemp for all his generous help in the preparation of this lecture.

REFERENCES

Allen, J.B. 1977 Two-dimensional cochlear fluid model: new results. *J. Acoust. Soc. Am.* **61,** 110-119.

Ashmore, J.F. 1987 A fast motile response in guineapig outer hair cells: the cellular basis of the cochlear amplifier. *J. Physiol.* **388**, 323-347.

Ashmore, J.F. 1989 Transducer motor coupling in cochlear outer hair cells. Pages 107-116 of Wilson & Kemp (1989).

Békésy, G. von 1960 *Experiments in Hearing*. McGraw-Hill.

Booker, J.R. & Bretherton, F.P. 1967 The critical layer for internal gravity waves in a shear flow. *J. Fluid Mech.* **27**, 513- 539.

Crawford, A.C. & Fettiplace, R 1980 The frequency selectivity of auditory nerve fibres and hair cells in the cochlea of the turtle. *J. Physiol.* **306**, 79-125.

Davis, H. and associates 1953 Acoustic trauma in the guineapig. *J. Acoust. Soc. Am.* **25**, 1180-1189.

Evans, E.F. & Wilson, J.P. 1973 The frequency selectivity of the cochlea. Pages 519-554 of Basic Mechanisms in Hearing (ed. A.R. Møller). Academic Press.

Forge, A. 1989 The lateral walls of inner and outer hair cells. Pages 29-35 of Wilson & Kemp (1989).

Gold. T. 1948 Hearing II. The physical basis of the action of the cochlea. *Proc. Roy. Soc. B.* **135**, 492-498.

Gold, T. & Pumphrey, R.J. 1948 Hearing I. The cochlea as a frequency analyzer. *Proc. Roy. Soc. B.* **135**, 462-491.

Holmes, M.H. 1982 A mathematical model of the dynamics of the inner ear. *J. Fluid Mech.* **116,** 59-75.

Hudspeth, A.J. 1985 The cellular basis of hearing: the biophysics of hair cells. *Science* **230**, 745-752.

Johnstone, B.M., Taylor, K.J. & Boyle, A.J. 1970 Mechanics of the guineapig cochlea. *J .Acoust. Soc. Am.* **47**, 504-509.

Kemp, D.T. 1978 Stimulated acoustic emissions from the human auditory system. *J. Acoust. Soc. Am.* **64**, 1386-1391.

Kemp, D.T. 1980 Towards a model for the origin of cochlear echoes. *Hearing Res.* **2**, 533-548.

Khanna, S.M. & Leonard, D.G.B. 1982 Basilar membrane tuning in the cat cochlea. *Science* **215**, 305-306.

Kiang, N.Y.S. & Moxon, E.C. 1974 Tails of tuning curves of auditory-nerve fibres. *J. Acoust. Soc. Am.* **55**, 620-630.

Kiang, N.Y.W. Watanabe, T., Thomas, E.C. & Clark, L.F. 1965 *Discharge Patterns of Single Fibres in the Cat's Auditory Nerve.* Massachusetts Institute of Technology Press.

Lighthill, M.J. 1962 Sound generated aerodynamically. *Proc. Roy. Soc. A* **267**, 147-82.

Lighthill, M.J. 1963 Jet noise. *Amer. Inst. Aeron. Astron. Journ.* **1**, 1507-1517.

Lighthill, M.J. 1972 The propagation of sound through moving fluids. *J. Sound Vib.* **24**, 471-492.

Lighthill, J. 1978a Acoustic streaming. *J. Sound Vib.* **61**, 391- 418.

Lighthill J. 1978b *Waves in Fluids.* Cambridge University Press.

Lighthill, J. 1981 Energy flow in the cochlea. *J. Fluid Mech.* **106**, 149-213.

Lighthill, J. 1983 Advantages from describing cochlear mechanics in terms of energy flow. Pages 63-71 of *Mechanics of Hearing* (ed. E. de Boer & M.A. Viergever). Delft University Press.

Lim, D.J., Hanamura, Y. & Ohashi, Y. 1989 Structural organization of the mammalian auditory hair cells in relation to micromechanics. Pages 3-10 of Wilson & Kemp (1989).

Lindsay, P.H. & Norman, D.A. 1977 *Human information processing.* Academic Press.

Neely, S.T. Norton, S.J. Gorga, M.P. & Jestendt, W. 1988 Latency of auditory brain-stem responses and otoacoustic emissions using tone-burst stimuli. *J. Acoust. Soc. Am.* **83**, 652-656.

Norton, S.J. & Neely, S.T. 1987 Tone-burst-evoked otoacoustic emissions from normal-hearing subjects. *J. Acoust. Soc. Am.* **81**, 1860-1872.

Peterson, L.C. & Bogert, B.P. 1950 A dynamical theory of the cochlea. *J. Acoust. Soc. Am.* **22**, 369-381.

Rayleigh, Lord 1896 *Theory of Sound*, 2nd edn. Macmillan. (see also the Dover reprint, 1945).

Rhode, W.S. 1971 Observation of the vibration of the basilar membrane in squirrel monkeys using the Mössbauer technique. *J. Acoust. Soc. Am.* **49**, 1218-1231

Rhode, W.S. & Robles, L. 1974 Evidence from Mössbauer experiments for nonlinear vibration in the cochlea. *J. Acoust. Soc. Am.* **55**, 588-596.

Robles, L. Ruggero, M.A. & Rich, N.C. 1986 Basilar membrane mechanics at the base of the chinchilla cochlea. I Input-output functions, tuning curves and response phases. *J. Acoust. Soc. Am.* **80**, 1364-1374.

Sellick, P.M., Patuzzi, R. & Johnstone, B.M. 1982 Measurement of basilar membrane motion in the guineapig using the Mössbauer technique. *J. Acoust. Soc. Am.* **72**, 131-141.

Steele, C.R. 1974 Behaviour of the basilar membrane with pure- tone excitation. *J. Acoust Soc. Am.* **55**, 148-162.

Steele, C.R. 1976 Cochlear mechanics. Pages 443-478 of *Handbook of Sensory Physiology*, vol. v/3 (Auditory System: Clinical and Special Topics, ed. W.C. Keidel & W.D. Neff). Springer.

Steele, C.R. & Taber, L.A. 1979 Comparison of 'WKB' and finite difference calculations for a two-dimensional cochlear model. *J. Acoust. Soc. Am.* **65**, 1001-1006.

Viergener, M.A. & Diependaal, R.J. 1983 Simultaneous amplitude and phase match of cochlear model calculations and basilar membrane vibration data. Pages 53-61 of *Mechanics of Hearing* (ed E. de Boer and M.A. Viergever). Delft University Press.

Voldrich, L. 1978 Mechanical properties of basilar membrane. *Acta Otolaryngol.* **86**, 331-335.

Voldrich, L. 1983 Experimental and topographical morphology in cochlea mechanics. Pages 163-167 of *Mechanics of Hearing* (ed. E. de Boer and M.A. Viergever). Delft University Press.

Wever, E.G. & Lawrence, M. 1954 *Physiological Acoustics* (see p.275). Princeton University Press.

Wilson, J.P. 1980 Evidence for a cochlear origin for acoustic re- emission, threshold fine-structure and tonal tinnitus. *Hearing Res.* **2**, 233-252.

Wilson, J.P. & Kemp, D.T. (eds.) 1989 *Cochlear Mechanisms - Structure, Function and Models.* Plenum.

Wit, H.P. & Ritsma, R.J. 1980 Evoked acoustical responses from the human ear: Some experimental results. *Hearing Res.* **2**, 253- 261.

Wit. H.P. & Ritsma, R.J. 1983 Two aspects of cochlear acoustic emissions: response latency and minimum stimulus energy. Pages 101-107 of *Mechanics of Hearing* (ed. E. de Boer and M.A. Viergever). Delft University Press.

Zenner, H.P., Reuter, G., Plinkert, P.K., Zimmermann, U. & Gitter, A. 1989 Outer hair cells possess acetylcholine receptors and produce motile responses in the organ of Corti. Pages 93-98 of Wilson & Kemp (1989).

Zweig, G. 1976 Basilar membrane motion. *Cold Spring Harbour Symposium on Quantitative Biology* **40**, 619-633.

Zwislocki, J. 1948 Theorie der Schneckenmechanik. *Acta Otolaryngol. Suppl.* **72.**

Zwislocki, J. 1986 Changes in cochlear frequency selectivity produced by tectorial-membrane manipulation. Pages 3-11 of *Auditory Frequency Selectivity* (ed. B.C.J. Moore & R.D. Patterson). Plenum.

Zwislocki, J. 1988 Tectorial membrane. I. Static mechanical properties *in vivo*. *Hearing Res.* **33**, 207-217.

THE RECENTLY RECOGNISED FAILURE OF PREDICTABILITY IN NEWTONIAN DYNAMICS*

Lecture Delivered on 4th December 1989
at National Aeronautical Laboratory, Bangalore

*Reproduced from: *Proc. R. Soc. Lond.*, A 407 (1986), 35-50.

INTRODUCTION

1. Newtonian Dynamics

(a) Giants and their shoulders

In the sixteenth century, students used to be sent quite young to the University of Copenhagen where, in fact, a particular thirteen-year-old student of philosophy woke up on the 21st August 1560 feeling very excited because the almanacks had forecast a solar eclipse for that date. Later in the day, when the eclipse did occur, exactly as predicted, the young Tycho Brahe received such a deep impression that he resolved to devote himself to the study of astronomy. He achieved some early fame in this field and then, in 1576, with the help of the Danish King Frederick II, he was enabled to build the admirably equipped observatory at Uraniborg. Here he accumulated an extraordinarily comprehensive body of observational data on the stars and the planets.

Much later, Tycho Brahe was, in effect, prevented from pursuing his work any further in Denmark. This happened in 1597 after Frederick's son, the less scientifically inclined King Christian IV, assumed power on attaining his majority. After some vicissitudes, however, Brahe was to acquire a new royal patron, the emperor Rudolph II; and, in 1599, to establish his equipment at the castle of Benatky in Bohemia, where he appointed as his assistant, the promising young astronomer Johann Kepler. Soon afterwards Brahe died but Kepler, as his successor and scientific executor, laboured conscientiously to publish in 1602 the magnificent body of data he had left behind. By 1627, furthermore, Kepler through his own observations had substantially augmented Brahe's data, as appeared in a new and very extensive publication which Kepler named the *Rudolphine Tables* in posthumous memory of the emperor.

Actually, Rudolph had been a rather exacting patron because of his obsession with horoscopes and astrology, which forced Kepler to devote massive efforts to fields secretly regarded by him as just what he called his 'means of subsistence'. Even so, he had been able, by 1609, to complete an extremely novel observational study of the motions of Mars; together with a full geometrical description of them, in the Copernican heliocentric frame of reference, as motions in a constant orbit in the shape of an ellipse with the Sun as focus. A decade later Kepler had published the *Epitome Astronomiae Copernicanae* (a rather more substantial work than the *Dialogo* which later got Galileo into some difficulties), and had there described in detail his most famous discovery: Kepler's three

empirical laws concerning planetary orbits. These laws, of the elliptical shapes of orbits, of the radius covering equal areas in equal times, and of the proportionality of the square of the orbital period to the cube of the major axis, were shown from the observations to be closely satisfied by the Earth and by the five then known planets; and, furthermore, by the four satellites of Jupiter which Galileo had recently discovered.

In beginning a lecture on Newtonian dynamics I have thought it appropriate to highlight some of those 'shoulders of giants' on which Newton recognized himself as standing; and, in particular, to recall the massive tasks of accumulating observational data, and of determining empirical laws that would describe them accurately, which made possible Newton's achievements. Nevertheless, I want to make a big distinction between any empirically based laws like those of Kepler, which may approximately summarize some large mass of data in a convenient and (perhaps) philosophically intriguing form, and physico-mathematical laws embodied in a system of general applicability such as Newton was the first to introduce. The distinctions are, of course, even sharper between empirical prediction of a solar eclipse like that of August 1560 (when the Sun's and Moon's observed movements would bring them into coincidence) and any predictability based upon physico-mathematical laws of motion.

(b) Newton's laws of motion

The modern world's practitioners of mechanics, for whom I do in some sense have the right to speak because I have just completed four years as President of the International Union of Theoretical and Applied Mechanics, feel pride in the knowledge that Mechanics was the first science for which a systematic framework of physico-mathematical laws was constructed, being expounded (of course) in the first book of Newton's *Principia;* a volume completed just over three hundred years ago and communicated to the Royal Society in the spring of 1686.

This framework is equally applicable to either of the two main parts of mechanics; this is, to statics, or to dynamics. These are concerned with how motion is prevented on the one hand, or governed on the other. Extending some ideas first introduced by Galileo, Newton adopted a strategy which may be summed up in the famous political maxim *Divide et impera:* divide and rule.

Thus, Newton's strategy requires first that we divide up or classify different types of *force*, representing different influences which in com-

bination may act to prevent motion, or else to govern it when it occurs. *Matter*, again, is to be thought of as divided up into particles of relatively small size, which are subject to forces of two main kinds: *external* forces (such as gravity, for example), along with those internal forces which may act between a pair of particles and which necessarily assume equal and opposite values.

Motion may be prevented in a system where, for each particle of the system, all the different forces acting on it add up to zero (add up, that is, in what we would now call a vectorial sense). Wherever the total force acting is non-zero, however, its value specifies the rate of change of *momentum* (mass times velocity) for the particle's motion; which, accordingly, it governs. For a large body such as a planet, the total momentum is its mass multiplied by the velocity of its mass centre, and the rate of change of this momentum is equal simply to the total *external* force, because all the (equal and opposite) internal forces acting between particles within it cancel out. Newton's inverse-square law of gravitation was called universal not just because he showed it to be the only law of force compatible with the three empirical laws of Kepler regarding motions of planets around the Sun (and of satellites around planets); but also because the same law accounted for gravitational phenomena at the surface of the Earth. In the case of an attracting body with spherical symmetry, the effect of attraction by all the particles making it up was proved equivalent to attraction by the whole mass concentrated at the centre; in such a way that another body moving solely in that field would describe an elliptic orbit with that centre as focus, and in accordance also with the other two laws of Kepler.

Newton recognized, at the same time, how dynamical effects of the Earth's rotation about its axis, taken in combination with his law of gravitation, accounted for the observed spheroidal shape of the Earth's surface, flattened at the poles. He appreciated, furthermore, how this departure from spherical symmetry made a certain *perturbation* in the gravitational force acting on the Moon which would partly explain the tendency of its orbit to process.

Post-Newtonian analyses of the solar system have been much concerned with developing such perturbation theories in far greater detail, with the object in particular of understanding all of those departures from the original simple empirical laws of Kepler which later astronomers were able to uncover in the course of their still more accurate observations. A particularly famous perturbation theory has been the three-body problem,

concerned with how the orbit of one body around another is perturbed by the influence of a third body. Many of the most important of the later developments in Newtonian mechanics (which, I should perhaps remind you, is the mechanics used for almost all practical engineering purposes) have been founded on mathematical methods first introduced in the context of perturbation theory; methods to which I must now briefly refer.

2. Mathematical Studies of Newtonian Dynamical Systems

(a) Developments in perturbation theory

Some big procedural advances in mathematical analysis, and in particular in the theory of differential equations, were needed for these subsequent major developments in Newtonian dynamics. Work done in the eighteenth century by the great Swiss mathematician Leonhard Euler initiated these advances, which were continued by Joseph Louis Lagrange. Between them they developed a method for calculating how different types of perturbations would effect a slow change in an orbit's so-called *elements*. These are a set of quantities defining the geometry of an orbital ellipse, and the orientation of its plane in space; quantities which under Kepler's laws would remain constant but which can vary slowly in response to perturbing forces of all kinds. Modern space scientists apply this method devised by Euler and Lagrange to relate observed variations in the orbital elements of artificial Earth satellites to the different types of perturbation causing them, and so to obtain precise estimates of strengths of different components in the perturbing forces; and this method gives us the best currently available values, in particular, for the departures of the Earth's real surface shape from the simple spheroidal first approximation obtained by Newton.

In the meantime, Lagrange himself had taken matters much farther in his *magnum opus*, the *Mécanique Analytique*, to obtain an absolutely general form for the system of differential equations satisfied by any mechanical system obeying Newton's laws. His own life's work, and that of his younger but equally gifted colleague Pierre Simon de Laplace, combined to yield clear explanations in terms of Newton's laws for practically all the peculiarities of orbital phenomena observed in the solar system. Laplace expounded this at length in his five-volume *magnum opus*, the *Mécanique Celeste*, and accompanied it with the brilliantly written single volume *Exposition du Système du Monde* (1796) which popularized the successes of Newtonian dynamics in accounting both for

the intricate behaviour of the solar system and for (broadly speaking) its stability. This work must have done much to spread into the general consciousness a belief in the complete predictability of systems based on Newton's laws of motion; a belief, as it were, in the determinism of the mechanical universe.

Here I have to pause, and to speak once again on behalf of the broad global fraternity of practitioners of mechanics. We are all deeply conscious today that the enthusiasm of our forebears for the marvellous achievements of Newtonian mechanics led them to make generalizations in this area of predictability which, indeed, we may have generally tended to believe before 1960, but which we now recognize were false. We collectively wish to apologize for having misled the general educated public by spreading ideas about the determinism of systems satisfying Newton's laws of motion that, after 1960, were to be proved incorrect. In this lecture, I am trying to make belated amends by explaining both the very different picture that we now discern, and the reasons for it having been uncovered so late.

The flowering of rigorous mathematical analysis in the nineteenth century, particularly among the French school of analysts, led to a situation where difficult questions about (say) the properties of solutions to a system of differential equations might at last be settled with complete certainty through the mathematical proofs of appropriate theorems. Now the equations of a very general dynamical system, given (as we have seen) in one quite useful form by Lagrange, had been transformed in the eighteen thirties by the brilliant Irish mathematician Sir William Rowan Hamilton into the still more convenient form of a set of simultaneous first-order differential equations; what we call the Hamiltonian formulation. Such a set of equations lent itself to study by rigorous mathematical analysis, which was attempted in particular at the turn of the century by Henri Poincaré (not to be confused with the distinguished statesman of the same surname). Poincaré (1892) was especially concerned with difficult cases of the three-body problem, including some of the interacting complexities of the Sun-Earth-Moon system which had continued to be hard to calculate.

Exact numerical computations of the solutions to such a Hamiltonian system of equations could be attempted, of course, only much later, in the post-1960 period, when the availability of powerful enough computers was certainly one of the main factors that altered our view of how such solutions behave. For Poincaré and for all of the pre-1960 analysts it was

necessary to base their analysis on perturbation theory; beginning, that is, from a very simple first approximation such as an orbit satisfying Kepler's laws, and then proceeding by taking into account perturbing effects (for example, of a third body) to a higher approximation, and then to successively higher and higher approximations, finally seeking to prove rigorously that the procedure converged to a well-determined limit. It was a programme that could be carried out very effectively in a wide range of cases. Nevertheless, in another important set of cases where there existed possibilities of approximate resonance between multiples of the different orbital periods of two interacting oscillations in a system, Poincaré's method of proof failed because of a difficulty that became known as the difficulty of small divisors, or resonant divisors. These were the denominators of terms in an infinite series in which every so often one of the terms had a very small denominator and this could often prevent the series from converging.

It was easy, of course, to assume that the failure of Poincaré's method of proof in a certain range of cases was accidental and that a regrouping of terms in the series, a sort of renormalization procedure such as achieved great success in other physical theories, would rectify the proofs so that they could be made rigorous. That type of approach was attempted for many years but it could never be shown to be applicable in full generality, and we now realize that this was due to a fundamentally new type of behaviour, which occurs in certain dynamical systems satisfying Newton's laws; a type which we call 'chaotic'. The rest of this lecture must be devoted to indicating how chaotic behaviour was discovered and to describing its nature [see Lichtenberg & Lieberman (1983) for a general survey].

(b) Chaotic behaviour

Three different new directions of research initiated the discovery of chaotic behaviour in simple systems satisfying Newton's laws during the early part of the 1960s; although, admittedly, it was only 20 years later that the remarkably widespread occurrence of chaotic behaviour had become clear; to such an extent that, for example, at the 1984 Congress of the International Union of Theoretical and Applied Mechanics the specialized topic which was highlighted most strongly in the Congress programme was that of chaotic behaviour.

The first of these three new directions of research arose directly out of Poincaré's work on nonlinear perturbation theory, related to equations in Hamiltonian form for isolated systems of constant energy. Work initiated

by the great Russian analyst Kolmogorov (1954), and pursued by his gifted colleague Arnold (1963), had been aimed first of all towards filling in the gaps in Poincaré's proofs; and, indeed, all this work along with independent studies by Moser (1962) in Germany and America demonstrated that, even in the neighbourhood of resonances, perturbations did assume a distinctly regular pattern in the vast majority of cases. Nevertheless, gaps in that regular pattern did exist; namely, very small ranges of initial conditions for which the motion assumed a form described as chaotic or stochastic (that is, random). It was regarded as interesting that equations of motion that included absolutely no random element should possess solutions which behaved in such a random way. Initially, however, the matter was seen as something of a curiosity because the highly complex proofs used in the perturbation theory were valid only when the perturbations were restricted to being sufficiently small and, in this case, the ranges of initial conditions for which chaotic or stochastic motions occurred were very limited indeed.

The second new direction of research utilized the powerful computers that were by then available to compute solutions not just in cases when the perturbations were small, but also for much larger perturbations. This work, carried out, for example, by Greene (1979) in the U.S.A. and by Chirikov (1979) in the U.S.S.R., demonstrated that as the strength of the perturbations continued to grow there was a sharp increase in the range of initial conditions for which solutions behaved stochastically. Finally, at a certain level of the perturbation strength, the authors observed what they called a transition to global stochasticity, with all solutions behaving chaotically. I shall describe later what this amounts to in detail, but in the meantime will note the obvious fact that the new data made a big increase in the importance to be attached to chaotic or stochastic solution.

These results on isolated systems of constant energy were of interest not only to astronomy but also to thermodynamics. The second law of thermodynamics envisages, of course, an increase in the randomness of motions experienced by an isolated system of molecules; that is, an increase in its entropy; but physicists had long supposed that large numbers of collisions between molecules were necessary to allow such randomization to occur. Now, with the wider understanding of how chaotic motions can develop, it is possible to see that collisions may not be essential. For example, the ionized gas between the Sun and the Earth with its extremely low density, producing an astronomically large mean free path between molecular collisions, may nevertheless in the presence

of magnetic fields experience phenomena that are possible only with an increase in entropy. One of these, which spacecraft have observed, is the so-called 'bow shock wave' where the solar wind of charged particles emanating from the Sun is abruptly slowed down where it first encounters the Earth's magnetosphere.

But that is a digression, which may on the other hand have provided a valuable reminder that Newtonian dynamics is applicable not only to systems of solid bodies but also to fluid systems, including ordinary gases and liquids. Ordinary gases and liquids, of course, are subject to the phenomenon called viscosity, which causes the mechanical energy in their shearing motions to be gradually dissipated into heat; precisely as a result of an entropy increase associated with normal molecular collisions on a submicroscopic scale. Yet even the damping of fluid motions by viscosity does not prevent perfectly regular fluid motions from becoming chaotic and this fact was first made precise over a century ago, in 1883, by Osborne Reynolds. He showed how the regular flow of fluid through a pipe suddenly becomes chaotic or turbulent when the force producing the motion becomes sufficiently large relative to the damping forces due to viscosity. He showed, furthermore, that this randomization has nothing to do with the random molecular movements occurring at submicroscopic scales; they, indeed, have a damping effect tending to reduce the trend towards turbulent motions; motions which themselves involve, rather, a chaotic pattern of fluid movement on a strictly macroscopic scale.

Thus, the specialists in dynamics of fluids, such as myself, have long been most fully conscious of the common tendency for regular or laminar motions of fluids to become chaotic or turbulent even though the motions in questions are subject to energy dissipation by the action of viscosity. On the other hand, fluids represent very complicated dynamical systems with an essentially unlimited number of degrees of freedom (each separate particle of fluid is separately free to be arbitrarily positioned relatively to all the other particles) and it had never been clear whether or not this was an essential pre-requisite for chaotic behaviour to develop.

Against that background it may be interesting to note that the third new direction of research which began in the early nineteen-sixties was concerned with some quite simple systems analogous to turbulence. These were dynamical systems with energy dissipation and just two or three degrees of freedom which, although forced in a perfectly regular way, responded in a completely chaotic way when a ratio of forcing effects: damping effects (a ratio similar to the Reynolds number introduced by

Reynolds) was sufficiently large. Initially, they were devised by some noted experts in dynamics of fluids, including the great atmospheric scientist E.N. Lorenz, in order to mimic as closely as possible the development of turbulence in fluid systems. Lorenz (1963) introduced the term 'strange attractor' to describe the type of randomized motion which inexorably tends to develop.

More recently, a very general theory of these strange attractors has been produced for such systems, which unlike the isolated energy-preserving systems studied by Poincaré and others, are subjected both to forcing and to damping (see chapter 7 of Lichtenberg & Lieberman, 1983). This theory suggests the steps by which regular motions develop into chaotic motions as some forcing:damping ratio changes. Often that takes place via an infinite sequence of so-called 'period-doubling bifurcations' which terminate, after just a finite change in that ratio, in a completely chaotic motion. Numerical computations have excellently confirmed these theories and demonstrated strong tendency for systems of this type also to develop chaotic motions.

Now at this point I might easily feel tempted to enlarge upon all the immense variety of different types of chaotic systems and of transitions to chaos[†]; and yet none of that would be relevant to the subject of this lecture. My objective in the time that remains to me is, rather, to focus upon certain properties which are common to all chaotic systems and which are relevant to the issue of predictability.

3. Limits To Predictability

(a) An illustrative, and relatively simple, chaotic system

These features connected with predictability that I shall describe from now on, then, are characteristic of absolutely all chaotic systems. Nevertheless, so as to fix our ideas, I shall concentrate attention first of all upon just one simple system; and, in fact, on the simplest of all those myriads of systems which have been proved to be chaotic.

This system is none other than the so-called spherical pendulum, consisting of a small weight attached to the lower end of a string of length l so that, when the top end is fixed, the weight is confined to a spherical surface of radius l. In small oscillations the weight has a certain natural period of sinusoidal oscillation given by the formula

$$T_0 = 2\pi\sqrt{(l/g)} \qquad \ldots (1)$$

[†]A recent paper, for example, classified eighteen different generic types of bifurcation (Stewart & Thompson, 1986).

Now this pendulum system has been shown to become chaotic if the top end is forced to move back and forth in a different sinusoidal oscillation with a period T just slightly greater than T_0. The maximum displacement in this forced motion is required to be small compared with l, and I shall take it to be l divided by 64. Then, John Miles (1984) established that the system is chaotic for a range of values of T around

$$1.00234T_0. \qquad \ldots (2)$$

provided that the damping afforded by, for example, air resistance does not exceed a certain (quite large) value. I shall give results for the case when one-tenth of the energy of the motion is dissipated by such damping during each period.

Now, for those of you who know about nonlinear oscillations I must make it clear that I am not simply saying that, in the case described, the naturally expected motion of the weight begins to be unstable. This behaviour, which might naturally be expected, would consist of a back-and-forth oscillation of the weight in the same plane as the forced oscillation of the upper end. I am not merely saying that this response begins to be unstable. That type of instability has long been very familiar in mechanics, as a matter of fact; and, in the present system, it occurs when T rises above $0.989\,T_0$. Then the planar oscillation becomes unstable, but there exist two simple non-planar oscillations that are stable, and the motion of the weight makes a gradual transition to either one or the other of these depending on its initial position. That behaviour is called a simple bifurcation.

In contrast with all that, the point about the present system is that an infinite number of further bifurcations take place as the forcing period increases before it reaches values around $1.00234T_0$ for which the system is chaotic. It is this infinite sequence of bifurcations which produces the completely arbitrary and randomized dependence on initial conditions which destroys predictability.

The second bifurcation occurs when the forcing period takes the value $0.99887\,T_0$. This leads the amplitudes of motion to start a slow regular fluctuation. Miles exhibits this on a planform plot which shows the successive positions of the weight every time the upper end of the string is at its position furthest to the right[†]. A simple planar or non-planar oscillation would be represented by a single point on this diagram. Instead

[†]In other words, this plot represents the amplitude of that part of the oscillation which is in phase with the forcing motion.

of that, the closed curve shown (Figure 1), around which the amplitude of fluctuation circulates every six and a bit periods, implies a spectacular variability in the appearance of the motion. This, actually, is the plot for a forcing period $0.99922T_0$. The spectrum of the motion indicates a fundamental frequency around 16% of the natural frequency of the pendulum.

The next plot (Figure 2), for a forcing period $1.00156T_0$, shows that a second bifurcation has occurred, with the appearance of subharmonics (that is, a halved fundamental frequency) and the amplitudes pursuing a double circulation. The third plot (Figure 3), for a forcing period $1.00222T_0$, demonstrates another halving of the fundamental frequency leading to still more subharmonies, and the amplitudes pursuing a quadruple circulation. Then the further bisections of fundamental frequency start to happen thick and fast: the next one is for a forcing period $T=1.00224T_0$ and an infinite number have occurred before T rises to the previously mentioned value $1.00234T_0$. Perhaps I should remark that for each of these motions where the plot lies mainly below the x-axis there is by symmetry another equally possible stable motion with the plot lying mainly above the x-axis.

Coming now to the plot in the chaotic case when the forcing period is $1.00234T_0$, we observe (Figure 4) a motion of a highly random nature *both* as regards the circulation of amplitudes (which has a form rather typical of chaotic motions) *and* as regards its noise-like spectrum. There is, furthermore, one special feature of this particular chaotic motion on which I would like to concentrate. This is that a typical circulation of amplitudes spends a longish interlude going around curves mainly below the x-axis, and then makes a transition to another longish interlude going around curves which are mainly above the x-axis (and represent approximate reflections of the former curves); and *then* it makes a transition back to another interlude mainly below, and then to another interlude mainly above, and then to another interlude mainly below and so on. The point is, that all these are longish interludes of unequal duration and that they are unpredictable in the sense that the sequence of successive durations of lower and upper interludes seems to vary not only randomly but discontinuously as a function of the initial conditions[†].

[†] For an experimental verification of Miles's theory, see Tritton (1986).

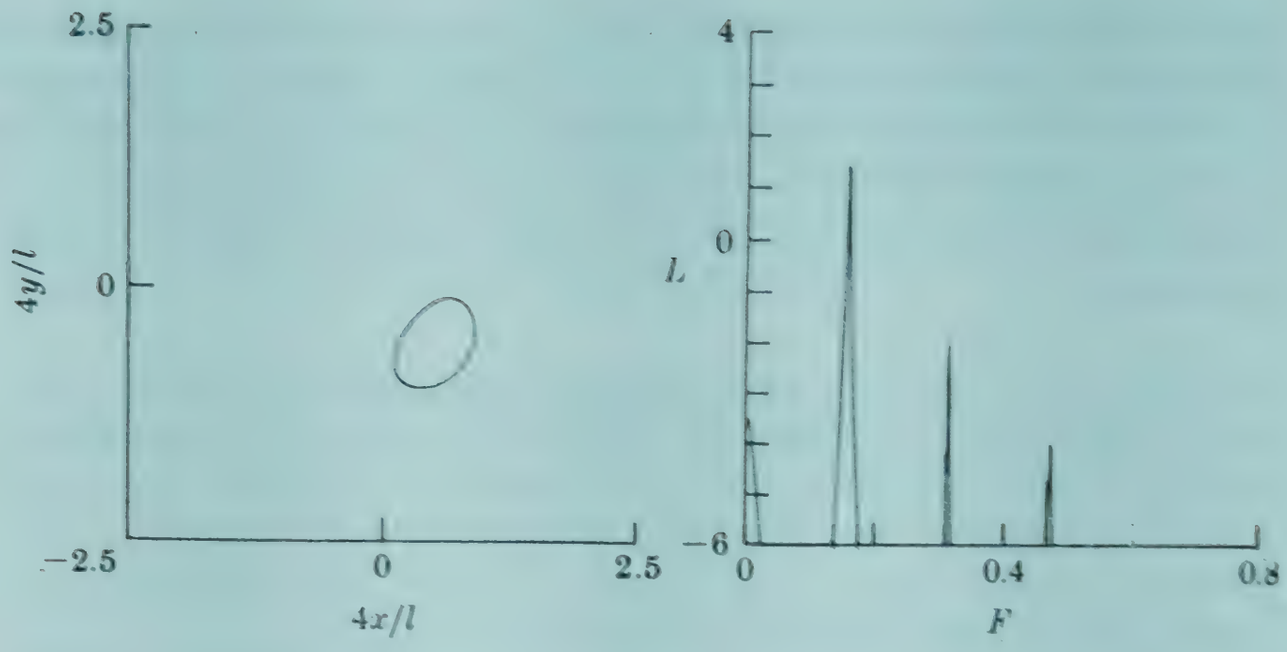


Fig. 1. Spherical pendulum forced by an oscillation of the point of support with amplitude l in the direction of the x-axis. The forcing period T is 0.99922 T_0. The left diagram plots the slow variation of the vector amplitude (x, y) of that part of the oscillation of the bob of the pendulum which is in phase with the forcing motion; the right diagram plots the frequency spectrum, with the logarithm L of spectral density shown as a function of the ratio F of frequency to natural frequency. (Figures 1-4 are reproduced, with permission, from Miles 1984.)

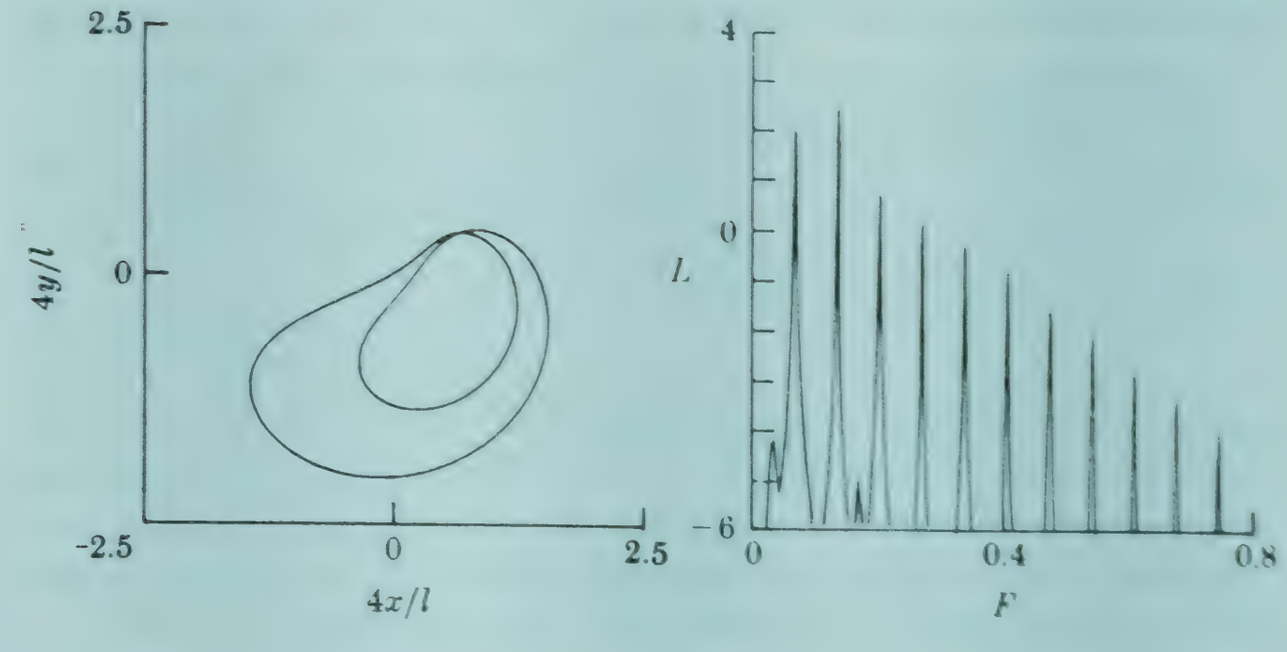


Fig. 2. Diagrams as in Figure 1 for the case when $T = 1.00156T_0$.

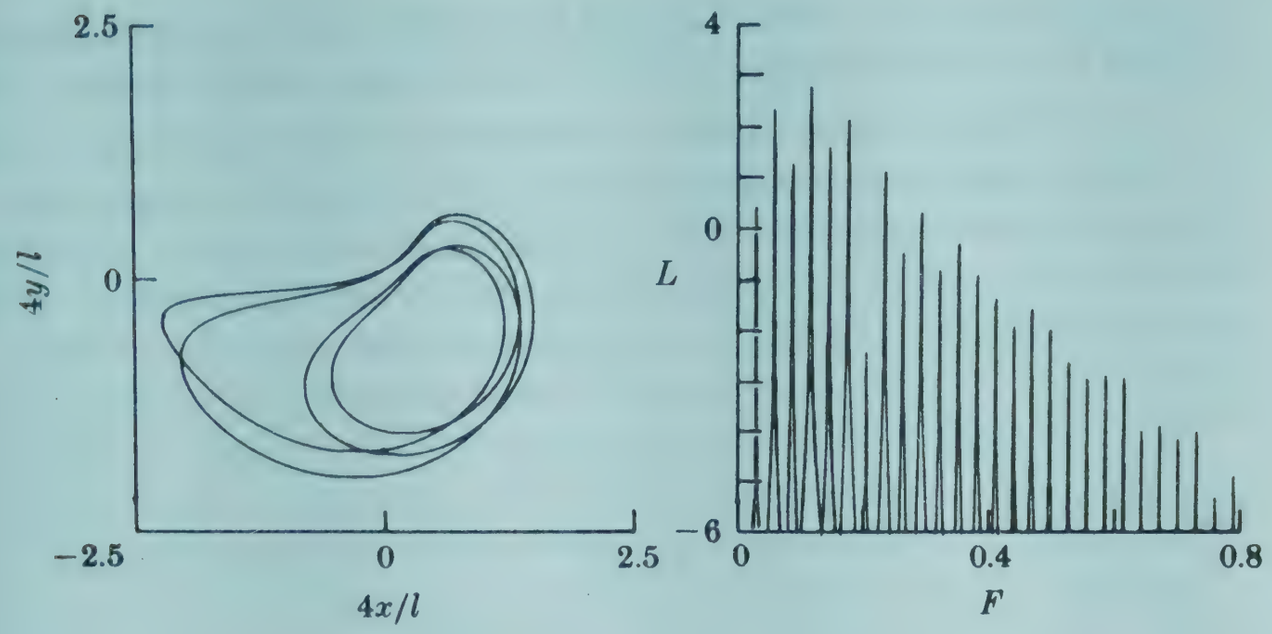


Fig. 3. Diagrams as in Figure 1 for the case when $T = 1.00222T_0$.

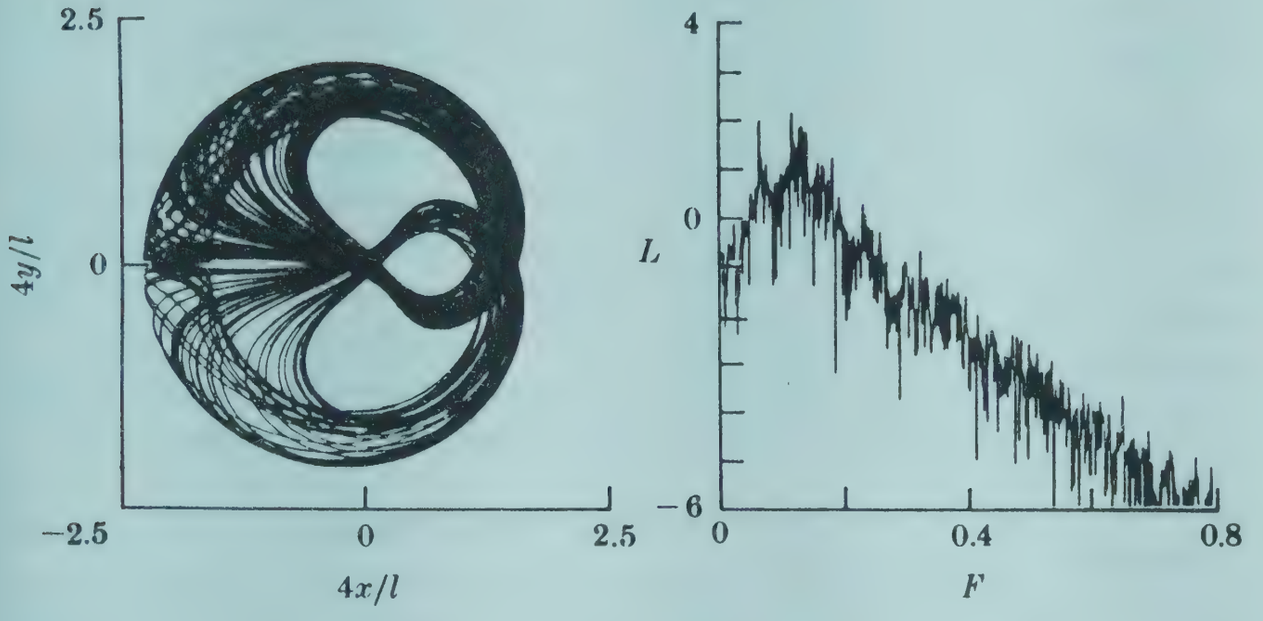


Fig. 4. Diagrams as in Figure 1 for the case when the forcing period T takes the value $T = 1.00234T_0$, within the range for which the response is chaotic.

Quite a lot of chaotic systems exhibit this particular *kind* of chaos with interludes of alternating polarity (as it were) and of randomly spaced durations. This is intriguing from the standpoint of geophysics because the magnetic field of the Earth, generated by motions within its liquid core, has been shown to have had a succession of interludes (all of many

tens of thousands of years) with such alternating polarity, and the durations of those interludes are distributed in a similarly random manner.

The question of predictability, of course, depends on how sensitively the motion that occurs depends on the initial position of the weight. This system governed by the laws of Newtonian dynamics would be predictable if, given the initial conditions to sufficiently high accuracy, the system's future could be approximately forecast. Let us take, for example, a hundredth of one per cent (that is 10^{-4}) as a reasonable level of accuracy in the initial conditions, so that we consider all possible initial conditions specified to four decimal places. Then the sequence of durations of interludes of alternating polarity is found to be altered discontinuously as we change from one initial condition to any of its nearest neighbours (in which there's a shift by just 1 in the *last decimal place*). The behaviour of the system, as calculated for two neighbouring initial positions of the weight, develops quite differently in ways that have already begun to diverge drastically beyond a certain time which we can call the predictability horizon. Changing the required accuracy of initial condition from (say) four decimal places to six decimal places produces only a modest increase in the predictability horizon.

I emphasize again that this is not just a case of solutions with different initial conditions beginning to diverge from one another as may happen in any region of instability. It is a case of completely randomized and discontinuous dependence of the solutions' later behaviours upon the initial conditions. This is the property that is specially characteristic of all known chaotic systems in their wide multiplicity and variety, and it is this property that I must now finally describe in more general terms.

(b) Limits to predictability for chaotic systems in general

With this end in view I obviously want to stop laying too much emphasis on any individual chaotic system; and, indeed the main reason for my selection of the one which I have just briefly outlined was that it is relatively easy to describe. Also, the fact that such a classic system as a pendulum (normally thought of as the epitome of regularity!) can in a certain range of circumstances, assume this chaotic property, which has been shown to be characteristic of such a wide range of more complicated systems, is seen by people in my discipline as a sensational discovery by Miles that deserves to be known still more widely. I remind you too that, the larger the perturbation from a simple regular motion, the *more* commonly chaotic behaviour occurs.

Naturally, the more complicated systems exhibit still more complicated forms of chaos than the one I have just outlined. In every case, however, certain common facts relevant to predictability persist.

Once again I remind you that, when a system is well defined, predictability depends upon the sensitivity of system behaviour to initial conditions. More precisely it is concerned with whether, given the initial conditions to whatever is the appropriate number of decimal places, the system's future may be approximately forecast.

Now, the common feature that is found to be characteristic of all chaotic systems is the existence of a predictability horizon. This is a time after which solutions with initial conditions that are 'nearest neighbours' to the accuracy of specification being used here become remote from one another; doing so, furthermore in a manner which varies in a discontinuous and randomized way in response to changes in just the last decimal place being used to describe the initial conditions.

You will want to ask, of course, what happens if we insist on *more* decimal places being used to describe the initial conditions. The answer is that predictability is changed rather little. In fact as the number of decimal places increases, the predictability horizon changes with it only *linearly*. This is because chaotic systems exhibit the property that neighbouring solutions diverge exponentially from one another (see, for example, p. 262 *et seq.* of Lichtenberg & Liebermann, 1983).

I must reiterate that this property takes a form quite different from the familiar type of exponential divergence of neighbouring solutions in an ordinary region of instability of some system. Very commonly, the fact that the system as a whole is of bounded total energy means that if solutions diverge from each other at one point they tend to come together again at another. With chaotic systems, however, this is not the case (even though they too, of course, have bounded total energy). In a chaotic system, nearby solutions are diverging from each other everywhere but they can do this without ever becoming infinitely far apart just because the chaotic behaviour is so random and diverse. Therefore, beyond the predictability horizon, neighbouring solutions have acquired randomly diverse characteristics.

I am inclined to insert a little parenthesis here and remark that actually, in certain areas of the dynamics of fluids, those concerned regularly carry out specific determinations of the predictability horizon. Meteorology is one of these, where we can employ any one of the presently available

global forecasting models and find out how different solutions diverge from one another if we vary the initial conditions in the last decimal place being used. The answer is not particularly sensitive to the number of decimal places used, and the theoretical predictability horizon so determined for global forecasting models is usually found to be a few weeks. This theoretical predictability horizon is firmly believed to be associated with the real chaotic behaviour of the system, rather than with the necessarily imperfect system of equations used to model it. Whatever the truth of that may be, any practical predictability horizons tend, of course, to be significantly less than those determined by the theoretical procedures I have described (because our knowledge is not nearly so perfect as those procedures assume).

Now I must come back to general statements about chaotic systems. I have pointed out that, within systems governed by Newtonian dynamics, it has been established that a fair proportion of them exhibit the chaotic property (and an *increasing* proportion of them as the degree of perturbation increases). For these, there is only a finite predictability horizon if initial conditions are prescribed to a given number of decimal places.

Now it may, of course, happen that some of you will wish to question the relevance of this conclusion by pointing out that, if the initial conditions were prescribed to an infinite number of decimal places, then the solution would be uniquely determined. In response to that, a mathematician might want in the first place to query and, indeed, to dispute whether the mathematician's own concept of the so-called 'real number' defined by an infinite sequence of decimal places has any relevance at all to measurements of what we habitually call 'the real world'. However, in the context of Newtonian dynamics we do not need to speculate about this.

Necessarily, we use Newtonian dynamics to describe the macroscopic behaviour of matter. This, however, can never be specified to more than a certain level of accuracy because of the well known Brownian motions associated with the continued vibrations of all the individual molecules; and, incidentally, we now recognize those vibrations of molecular systems as having the type of random behaviour which they are known to exhibit (a type of random behaviour usually called 'ergodic') just because of the propensity towards chaos in mechanical systems satisfying Newton's laws. Initial conditions can never be specified, then, beyond a certain level of precision, and this implies a predictability horizon.

I feel fully justified, therefore, in repeating that systems subject to the laws of Newtonian dynamics include a substantial proportion of systems that are chaotic; and that, for these latter systems, there is no predictability beyond a finite predictability horizon. We are able to come to this conclusion without ever having to mention quantum mechanics or Heisenberg's uncertainty principle. A fundamental uncertainty about the future is there, indeed, even on the supposedly solid basis of the good old laws of motion of Newton, which effectively *are* the laws of motion satisfied by all macroscopic systems[†]. I have ventured to feel that this conclusion would be of interest to a rather general audience assembled here in the distinguished scientific city of Bangalore.

REFERENCES

Arnold, V.I. 1963 *Soviet Math* **3**, 136.

Chirikov, B.V. 1979 *Physics Rep.* **52**, 265.

Greene, J.1979 *J. Math. Phys.* **20**, 1183.

Kolmogorov, A.N. 1954 *Dokl. Akad. Nauk SSSR* **98**,527.

Lichtenber, A.J. & Lieberman, M.A. 1983 *Regular and Stochastic Motions*. New York: Springer.

Lorenz, E.N. 1963 *J. atmos. Sci.* **20**, 130.

Miles, J. 1984 *Physica* **D11**, 309.

Moser, J. 1962 *Nachr. Akad. Wiss. Göttingen, Math. Phys. Klasse p.*1.

Poincaré, H. 1892 *Les Methodes Nouvelles de la Mécanique Celeste*. Paris: Gauthier-Villars.

Stewart, H.B. & Thompson, J.M.T. 1986 Towards a classifications of generic bifurcations in dissipative dynamical systems. In *Dynamics and Stability of Systems*, vol. 1. Oxford University Press.

Tritton, D.J. 1986 *Eur. J. Phys.* **7**, 162.

[†]That is, systems for which phenomena on the extreme sub- microscopic scales specified by Planck's constant are negligible.

BIOFLUIDDYNAMICS OF BALISTIFORM AND GYMNOTIFORM LOCOMOTION*

PART 1. BIOLOGICAL BACKGROUND, AND ANALYSIS BY ELONGATED-BODY THEORY

(In Collaboration with Robert Blake†)

Lecture Delivered on 5th December 1989
at Central Leather Research Institute, Madras

*Reproduced from *J. Fluid Mech.*, 212(1990), pp.183-207. Parts 2, 3 and 4 represent mathematical appendices and will appear in vol. 213 of the same journal.

†Department of Zoology, University of British Columbia, Vancouver, V6T 1W5, Canada.

1. INTRODUCTION

The two-fold challenges of *Mathematical Biofluiddynamics* as described in a book with that title (Lighthill, 1975; hereinafter referred to as *MB*) are derived from the problems of mathematically analysing complicated three-dimensional unsteady flows 'energized by the working of an animal's motile external or internal surfaces, parts of which are highly flexible' (*MB*, p.1) while recalling that 'the analyses required must be carefully defined and interpreted in a full and intricate biological context' (*MB*, p. 2). These problems are well illustrated by the study of fish locomotion, set in its biological context in *MB*, pp. 11-44 as well as in more recent monographs (Blake, 1983a; Webb & Weihs, 1983), and shown (*MB*, pp. 103-114) as demanding from the mathematician a large amplitude analysis of the three-dimensional unsteady fluid motions.

For the quantitative study of how in a wide range of fish species body undulations of large amplitude are used for the efficient execution both of forward swimming and of fast starts and turns, there remains in the 1980s (just as there was in the 1970s) no realistic option of applying high-Reynolds-number computational fluiddynamics to calculate successive flow fields in a three-dimensional fluid region whose shape is making large changes with time. Accordingly, the analysis by 'large-amplitude elongated- body theory' (*MB*, pp. 132-137) which can be readily applied in a wide range of cases, eventhough it describes the motions to only a crude approximation, must be our main source of comparative information on when such body undulations can achieve good hydromechanical efficiency*.

The evident need to extend elongated-body theory to modes of swimming in which the body remains rigid and propulsion is achieved by undulations propagated along a fin alone — or a pair of fins — was noted in a list of problems requiring further study (under the heading (a) in *MB*, p. 42, which refers back to descriptive material on p. 37) but has not hitherto been attempted. Yet there are biologically important unsolved questions, set out in detail in §2 below, which call for just this type of analysis.

* We should however make it clear here that we are by no means discounting the possibility that in the future an approximate treatment on somewhat different lines might be shown to be even more accurate for the description of certain types of fish locomotion, including cases of balistiform swimming.

In particular, the family Balistidae of rigid-bodied fishes (popularly called trigger-fishes) move unexpectedly fast by means of 'balistiform' propulsion; that is, propulsion by undulations propagated simultaneously along a posterior pair of highly flexible dorsal and anal fins. Analysis in this paper is comprehensively applied to understanding the effectiveness of balistiform propulsion in this and other families of fishes, and is partly extended to 'gymnotiform' propulsion (where, as in the electric eel *Gymnotus*, propulsion may occur through undulations propagated along a single ventral fin). This analysis is a direct (although, as we shall see, not a completely straightforward) extension of the large-amplitude elongated-body theory used to study fish locomotion by means of body flexure, of which we now briefly recall the fundamental hydromechanical ideas.

Although this latter theory can be recognized as having developed from ideas of long standing used in 'slender-body theory', the alternative designation 'elongated-body theory' has been preferred in fish biofluiddynamics, partly because of its greater intelligibility to biologists, but mainly to highlight the special characteristics of a theory requiring careful treatment of the fluid transition between the elongated region occupied by the fish's propulsive surfaces and the posterior vortex wake. Elongated-body theory studies the rate of change of fluid momentum within a volume $\mathcal{V}$ bounded by a control surface which includes a posterior plane Π, separating the regions of attached vorticity around the body and fins from the region of shed vorticity in the wake. For a complete 'budget' of the momentum balance it is necessary to take into account momentum transfer by convection across the plane Π and also by means of the overall pressure force acting across Π.

In all other respects, the nature of the analysis is similar to that in slender-body theory. The fluid momentum per unit length of fish around each cross-section is taken to be that which would be found in two-dimensional irrotational flow generated by the instantaneous motion of that cross-section resolved at right angles to the backbone. Because, in addition to the posterior plane Π, the control surface bounding the volume $\mathcal{V}$ includes two parallel horizontal planes[†] above and below the plane of the fish's swimming movements (here taken as horizontal), the fluid momentum associated with that cross-sectional motion is well-defined. Indeed,

[†]The 'budget' of horizontal momentum within $\mathcal{V}$ cannot, of course, include any term due to the pressure forces acting across these horizontal planes.

when the complete cross-section moves as a whole with velocity w, that momentum takes the classical form mw, where m is the cross-sectional added mass per unit length. Furthermore, the area integral of the pressure in the fluid region between the planes (a quantity needed, as we have seen to specify the overall pressure force acting across Π) takes the positive value $\frac{1}{2} mw^2$.

We may pause to compare this elongated-body theory with the types of slender body theory that are much used to describe microorganism locomotion at low Reynolds numbers, with 'stokeslet' singularities distributed along the centreline of a waving flagellum. Those theories are purely linear and neglect all inertial effects; but are subject to complications resulting from the long-range character of the stokeslet velocity field, which causes substantial interactions between motions at different cross-sections. By contrast, the elongated-body theory is nonlinear and largely dominated by inertial effects (as in the form of the convective transfer of momentum across Π); but the velocity fields associated with cross-sectional motions possess the relatively short-range character associated with dipole far fields, so that interactions between motions at different cross-sections are greatly reduced.

It will be evident that, if and when elongated-body theory is to be extended to balistiform swimming (for example), the novel analyses required must include certain new two-dimensional calculations. First of all, the fluid momentum per unit length must be derived for motions of a fish cross-section where the body element is stationary and only the fins are moving. Secondly, it is necessary to calculate the area integral of the fluid pressures associated with such motions in order to derive the overall pressure force acting across the posterior plane Π .

Out of this pair of new two-dimensional calculations, we are led by the existing elongated-body theory to expect the first to make a bigger contribution to thrust than the second. Accordingly, special attention has been given in our studies (see especially Part 3 of this paper, Lighthill, 1990b) to calculating the fluid momentum associated with the movement of fins attached to a rigid-body cross- section.

The importance of this line of research became still clearer after preliminary calculation for a simple model of balistiform swimming, given in §4 below, showed a large enhancement of fluid momentum. Specifically, the fluid momentum produced by movements of dorsal and anal fins attached to a deep rigid-body cross- section was found to be from

3 to 4 times greater than the same fin movements would generate on their own.

By contrast, an interesting calculation of the area integral of pressure, given in Part 2 of this paper (Lighthill, 1990a) for the same model, demonstrated that this does still make a positive contribution to thrust (as in the well established elongated-body theory), but a contribution which is in no way enhanced relative to that generated by the fin movements on their own. Accordingly, it is even less important in comparison with the contribution from fluid momentum than in the case when cross-sections move as a whole.

The striking nature of the conclusions regarding enhancement of fluid momentum made it important to check whether such enhancement occurs for accurately realistic models of balistiform swimming, in which the body cross-section is given the form of an ellipse with arbitrary axis ratio (the earlier model represents a limiting case of small axis ratio). In Part 3 appropriate conformal mappings are used to carry out this calculation, and to extend it also to gymnotiform swimming (involving motions of just a single fin). In every instance a similarly substantial enhancement of fluid momentum is found.

Nevertheless we suggest that this is just one of two important effects contributing to the 'uncannily' high-speed swimming movements observed in trigger-fishes. Certainly, the large-amplitude undulations of modest-sized dorsal and anal fins carry, as we have seen, enhanced momentum which generates an enhanced component of thrust. It is just as important, however, that the body drag, which the fin thrust must overcome, is reduced to the value associated with the uniform rigid forward motion of the body through the water; this contrasts with the value much greater (by a factor of 3 to 4) that has long been accepted (*MB*, pp. 114-116; see also Blake, 1983a, pp. 98-101) as accompanying body undulations. The thrust enhancement and drag reduction are both by a factor of around 3 or a little more and, taken together, are sufficient to explain a similar speed enhancement for balistiform swimming.

Our conclusions are important also for gymnotiform swimming, which they indicate as conferring substantial advantages through reduction of the energy cost of transport. This idea tends to reinforce an earlier suggestion (Blake, 1983b) that evolution of gymnotiform swimming was not primarily related to the special needs of electric fishes; a suggestion consistent, indeed, with the use of gymnotiform swimming by a species,

Xenomystics nigri, from a family Notopteridae (the knife fishes) lacking any electrical organs.

From the general standpoint of theoretical fluid mechanics, the careful formulation in §3 of an extension of large-amplitude elongated-body theory to fish propulsion by fin undulations is of some conceptual interest; whereas the detailed calculations of fluid momentum in Part 3, although producing interesting results, involve methodologically a straightforward application of conformal-mapping techniques. Finally, the calculation of pressure in Part 2 for irrotational flows produced by motions of a flexible boundary proves (alongwith the deduction of its area integral) to involve more subtle analysis than might be expected, essentially because in the unsteady form of Bernoulli's equation the $\partial\phi/\partial t$ term (time-derivative of the velocity potential) is not simple to obtain when the shape of the boundary is itself varying with time.

It is a pleasure to dedicate to Professor G.K. Batchelor on his 70th birthday a paper which in its Part 2 includes material on fluid pressure distributions produced by body flexure that bears some relation to his fundamental interests in mechanics of fluids as well as to his outstanding text on the subject (Batchelor, 1967), while in §3 below it provides a new application of those 'slender-body theory' ideas in which he has long taken a profound interest.

2. Biological background

Modes of propulsion of aquatic animals are of particular hydromechanical interest because about 10^9 years of animal evolution in an aqueous environment, by preferential retention of specific variations that increase ability to survive and produce fertile offspring, have inevitably produced rather refined means of generating fast movements at low energy cost, which merit study. This is mainly because at every point in the food chain, or network of predator-prey relationships, capability of movements that are fast relative to those of predator or prey, or to the distances that must be traversed before prey is found, is one out of the many important factors influencing such ability to survive.

Some of the other factors, to be sure, may in certain ecological situations be of overriding importance, so that evolution proceeds in a direction — for example, development of protective armour — that may reduce mobility although net chance of survival has improved. Defensive armour has been developed at the expense of speed within three important

orders of fishes. (Tetraodontiformes, Syngnathiformes, Zeiformes[†]) where the resulting loss of body flexibility has been such as, in general, to preclude propulsion by body undulations. The principal alternative modes of propulsion found in such fishes are reviewed by Blake (1983c) as well as in Blake (1983a, chap. 6).

Amongst all of these modes of fish propulsion, we discuss here just one, which can perhaps be regarded as a further adaptation, in fishes that had lost body flexibility, so as to allow them nevertheless to achieve forward motion at low energy cost through a combination of good propulsive efficiency with avoidance of drag enhancement. This is the balistiform mode of swimming, associated with undulatory movements made by both of the median (dorsal and anal) fins. Here it is discussed in the first place for rigid-bodied fishes, but afterwards some cases of flexible-bodied fishes which are able under a variety of different circumstances to use either anguilliform body undulations or balistiform median-fin undulations are described and interpreted. The section is concluded by a discussion of the gymnotiform mode of swimming, which (this time normally in flexible-bodied fishes) involves undulations in just a single median fin.

The balistiform mode of swimming, as a means of economically achieving more rapid locomotion than might seem compatible with a nearly rigid body, is found above all in the sub-order Balistoidei of the order Tetraodontiformes; and, particularly, in the family Balistidae (trigger-fishes, file-fishes and leather-jackets).

Figure 1 shows a range of species from the Balistoidei, all of which achieve forward locomotion primarily through simultaneous movement of posterior dorsal and anal fins. They are divided into two main groups: those in genera like *Balistes*, *Abalistes*, *Monacanthus*, *Oxymonacanthus*, and *Aluterus* where the undulation passing backwards along both fins has a wavelength considerably less than the lengths of those fins; and those in genera like *Pseudobalistes*, *Odonus*, *Canthidermis* and *Xanthichthys* where the oscillatory motion of the fins, if it were to be described as an undulation, would need to be identified as one that fails to include any entire wavelength.

Thus, any comparison between the two types of balistiform swimming must be rather similar to the comparison between body undulations in the anguilliform and carangiform modes (Gray, 1968, pp. 19-38; *MB*, pp.

[†]Designated in *MB* by their older names Plectognathi, Solenichthyes, Zeomorphi.

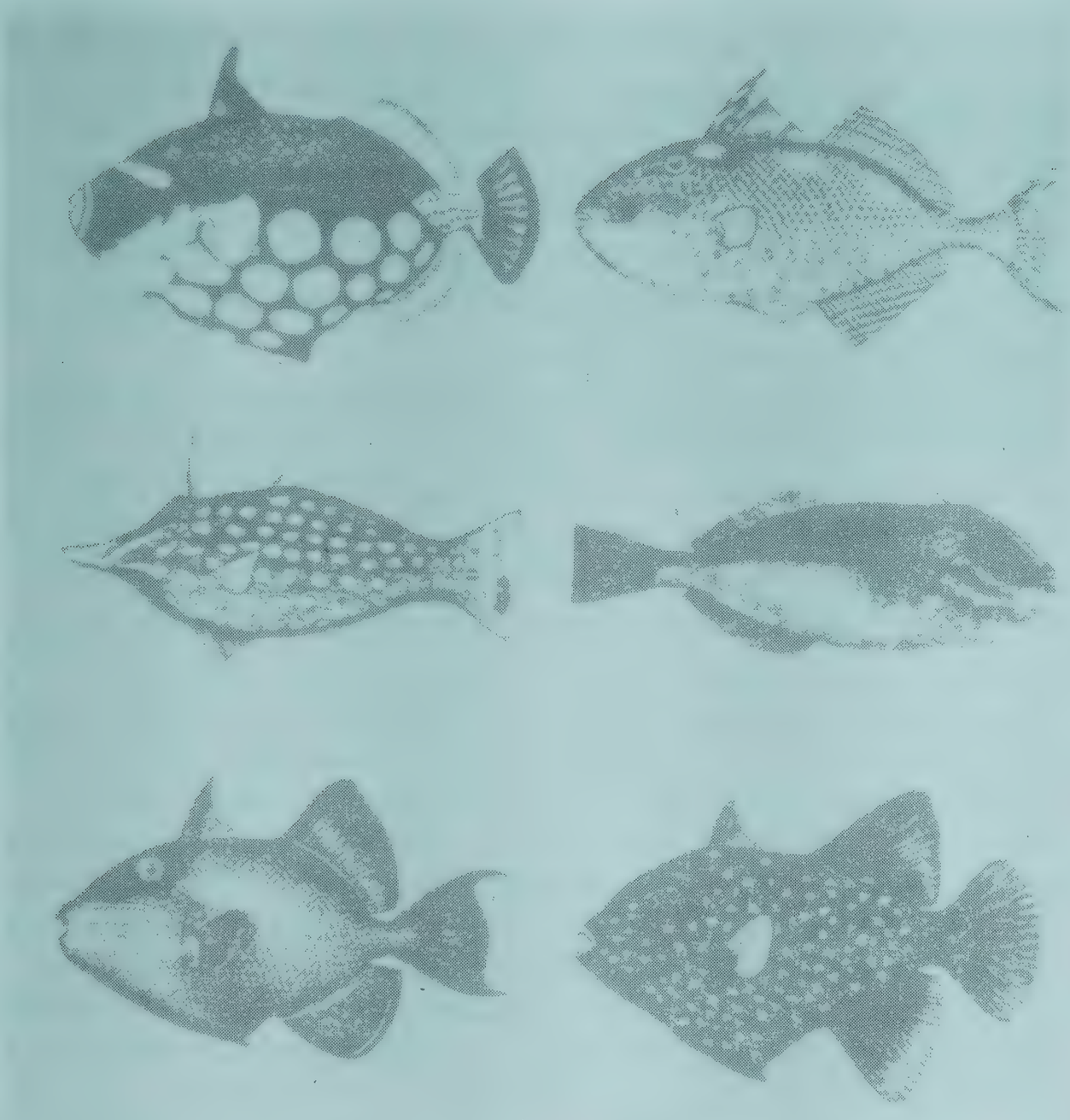


Fig. 1. Some fish species from the sub-order Balistoidei which use balistiform locomotion, either in the undulatory mode as with (*a*) *Balistes conspicillum*, (*b*) *Abalistes stellaris*, (*c*) *Oxymonacanthus longirostris* and (*d*) *Aluterus monoceros*, or in the oscillatory mode as with (*e*) *Pseudobalistes fuscus* and (*f*) *Canthidermis rotundatus*. Photographs in Figures 1 and 2 are taken from Marshall (1964), except for Figure 2(*b*) which comes from Marshall (1965).

19-26). The second type may indeed be seen as involving not so much a wave as a lateral oscillation of each fin as a whole, with a twist at each extreme of the oscillation so that the fin's side-to-side movement includes always a backward-facing component.

For all of the fishes in Figure 1 the assumptions of elongated-body theory can represent only a crude approximation, but the theory neverthe-

less gives us some useful guidance on the forces exerted by the fin motions. Broadly, it suggests that, although the two types of balistiform swimming exert rather similar propulsive forces, nevertheless the fully undulatory mode is a more economical way of achieving forward movement. Above all, it keeps to a minimum any fluctuating sideforce producing lateral oscillations of the body (oscillations described as 'recoil' in *MB*, p. 87), which is important because it avoids enhancement of the body drag which fin thrust has to overcome.

The theory also indicates that the relatively low ratio of fin depth to body depth characteristic of species utilizing the fully undulatory mode of balistiform swimming does not necessary lead to reduced propulsive forces. It suggests, in fact, that the force associated with fin movements of a given magnitude is substantially enhanced when this ratio is small.

Among the other two sub-orders of the Tetraodontiformes the truly undulatory mode of balistiform swimming is not found, although several species use balistiform swimming of the oscillatory type, normally as one of a variety of different available methods of locomotion. In the sub-order Tetraodontoidei the majority of species are somewhat sluggish swimmers; on the other hand, fishes of the genus *Tetraodon* (see Figure 2) are able to use oscillation of the dorsal and anal fins for reasonably fast forward locomotion. The sub-order Ostraciontoidei, on the other hand, is best known for the so-called ostraciiform mode of swimming, involving a simple rotary oscillation of the caudal fin about the base of the caudal peduncle, which is not considered to be of good hydromechanical efficiency (Blake, 1981).

Within the other two main orders of fishes where development of protective armour has reduced body flexibility, it is above all the Zeiformes (John Dories, boar fishes, etc.) which have adopted the truly undulatory mode of balistiform swimming. For example, the boar-fish *Capros aper* (Figure 2) is typical of fishes in this order in its use of undulations propagated along opposed soft dorsal and anal fins to propel itself forward. The qualitative conclusions of this paper may be broadly relevant to the understanding of balistiform swimming in zeiform fishes; where, however, an unusually low ratio of overall length to body depth makes the use of elongated-body theory much less appropriate than in the case of the Balistoidei.

On the other hand, within the order Syngnathiformes (the tube-mouthed fishes, from sea-horses and pipe-fishes to cornet-fishes and

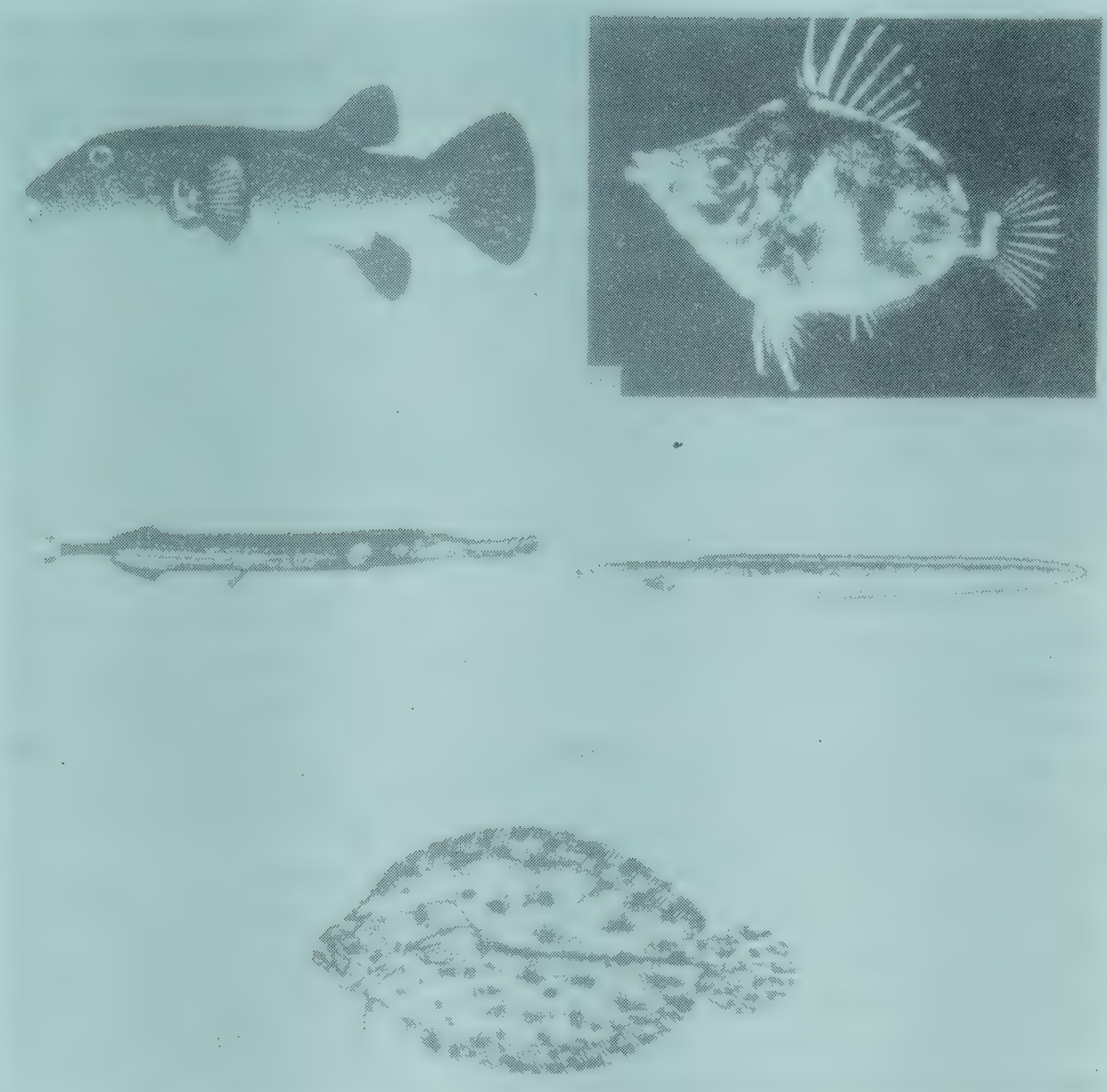


Fig. 2. Some other fish species able to use balistiform locomotion: (*a*) *Tetraodon stellatus*, (*b*) *Capros aper*, (*c*) *Aulostomus chinensis*, (*d*) *Conger labiatus* and (*e*) *Pseudorhombus arsius*

trumpet-fishes), it is mainly the family Aulostomidae (trumpet-fishes) that uses the truly undulatory mode of balistiform swimming. A fish like *Aulostomus chinensis*, also shown in Figure 2, adopts this as its regular mode of swimming, which, however, it can supplement with caudal-fin motions when hard pressed.

Among the flexible-bodied fishes, there are several with long dorsal and anal fins which are able to use the truly undulatory mode of balistiform swimming as an alternative low-cost means of progression when they do not need the higher speeds available with angulliform body undulations.

This is particularly characteristic of the order Anguilliformes† (the eels), as illustrated in Figure 2 by the conger-eel *Conger labiatus*.

The same tendency is characteristic of the Pleuronectiformes† (flat-fishes) which, it must be emphasized, adopt the characteristically anguilliform mode of swimming in their very early stages of life, after which a flat-fish turns over onto one side while the eye on that side moves round to the other. It retains the anguilliform mode of swimming for faster movements and also, as in Anguilliformes, the undulatory mode of balistiform swimming for relatively slower movements; but the motions of body and/or fins after it has turned over are, of course, vertical. The median fins in a flat-fish are usually referred to as 'marginal fins' (see Figure 2 for the case of a flounder *Pseudorhombus arsius*) and, in the balistiform mode of swimming, their undulatory motions are similar to those observed in Balistoidei except for being turned through 90°.

Although the principal analysis in this paper is for balistiform swimming, we conclude this section by indicating the biological background to our briefer analysis of gymnotiform swimming. This is characteristic of the 'electric eels' of the two families Gymnotidae and Electrophoridae in the order Cypriniformes. The electric eels have flexible bodies and are capable of anguilliform body undulations. For the most part, however, they choose to keep the body rigid and to propel it forward by means of undulations propagated along their one and only median fin: the ventral fin. Figure 3 gives synchronous side and bottom views of gymnotiform swimming in two gymnotid species (*Gymnotus carapo* and *Gymnorhamphichthys hypostomus*); it has been argued (Lissmann, 1958) that this mode of swimming facilitates effective utilization of the fishes' electrical organs.

On the other hand, Blake (1983b) has drawn attention to the fact that an unrelated species of fish without any electric organs (*Xenomystis nigri* from the family Notopteridae in the quite different order Osteoglossiformes) uses a similar type of gymnotiform swimming, also shown in figure 3. He uses this observation along with others (to the effect that certain mormyroid fishes are able to use electric organs effectively while swimming by means of body motions in the carangiform mode) to suggest that other advantages besides those specialized to electric fishes may have contributed to the evolution of gymnotiform swimming. Briefly, these

---

†The older names Apodes and Heterosomata were used in *MB* for the orders Anguilliformes and Pleuronectiformes.

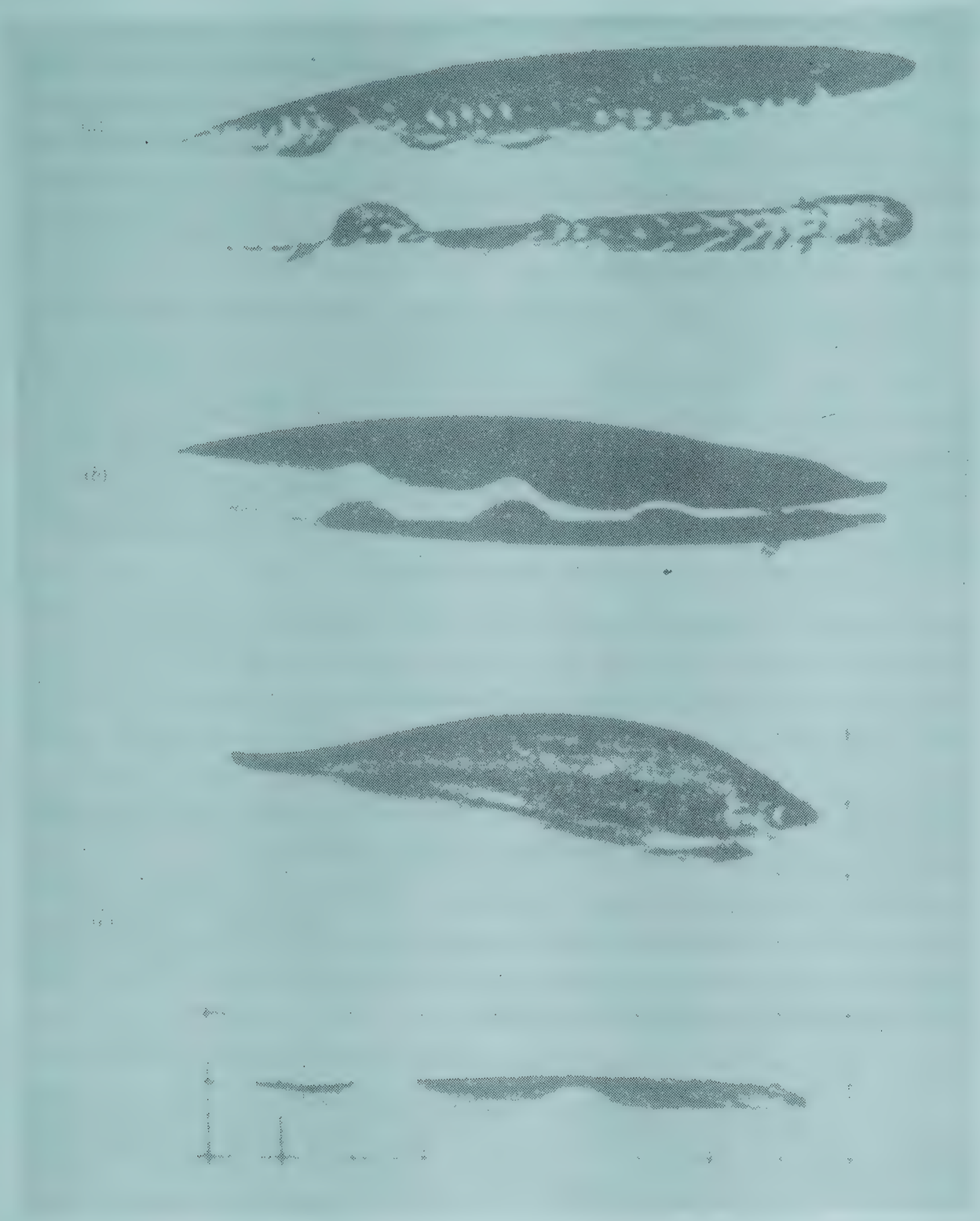


Fig. 3. Synchronous side and bottom views of gymnotiform swimming in two gymnotid species: (*a*) *Gymnotus carapo* and (*b*) *Gymnorhamphichythys hypostomus*, and in one notopterid species: (*c*) *Xenomystis nigri*.

include the avoidance of fluctuating sideforce and hence of any increase of drag (through body oscillations) over its rigid-body value, together with an enhancement of fin thrust as indicated by calculations in Part 3 (from which a preliminary result had been noted already by Blake, 1983b).

3. Extension of elongated-body theory

In this section we first recall the results (*MB*, pp. 132-134) of elongated-body theory for fish swimming by body undulations in a horizontal plane where each fish cross-section is subjected only to horizontal translation and to rotation about a vertical axis. In vector notation, these results derive the horizontal force $\boldsymbol{F}$ with which water pressures act on the fish from the following principles:

(a) The fish acts on the water with an equal and opposite force ($-\boldsymbol{F}$);

(b) If we consider a region $\mathcal{V}$ of the fluid bounded internally by the fish and externally by a certain control surface, then the rate of change of horizontal water momentum in $\mathcal{V}$ can be equated to sum of the forces acting on it (namely, $-\boldsymbol{F}$ plus the sum of the horizontal pressure forces acting across the control surface) minus the rate of transport of horizontal momentum out of $\mathcal{V}$ across the control surface;

(c) The control surface includes two horizontal planes at equal distance above and below the plane of the body's undulations, and this distance is taken sufficiently large so that rate of transport of momentum across the planes (a quantity of second order in the disturbances) is negligibly small; while, because the planes are horizontal, the pressure force on each — although it may not be negligibly small (since the pressure is of first order in the disturbances) — has zero horizontal component;

(d) The control surface includes also a vertical plane Π, attached rigidly to a posterior cross-section of the fish (and therefore undergoing the same translations and rotations as that posterior cross-section), so that it remains always perpendicular to the backbone;

(e) Shed vorticity is found entirely behind the posterior plane Π;

(f) At each instant, therefore, the horizontal momentum of the fluid within $\mathcal{V}$ (that is, anterior to Π) can be related to the irrotational flows generated by the motion of body cross-sections;

(g) The momentum per unit length of body at every cross- section can be approximated by the value $\boldsymbol{M}$ that it would take in a two-dimensional flow associated with the motion of that cross-section;

(h) Similarly, the pressure force acting on $\mathcal{V}$ across Π (which, as explained in (b) above, contributes to the expression for the propulsive force $\boldsymbol{F}$) can be approximated by the value $\boldsymbol{P}$ that it would take in a two-dimensional irrotational flow in the plane Π associated with the motion of the posterior cross-section.

To this list of principles underlying the theory, we may add the facts that (i) the momentum per unit length of body referred to in (g) — that is, the horizontal momentum $\boldsymbol{M}$ of the fluid between the parallel horizontal planes — is directed perpendicular to the backbone, in the plane of the cross-section; (ii) its magnitude can be expressed as mw in terms of the added mass m of the cross-section and its velocity component w perpendicular to the backbone; (iii) the magnitude of the pressure force $\boldsymbol{P}$ referred to in (h) can be calculated as the area integral of the pressure in the region between the two parallel planes for the two-dimensional flow produced by the cross-section's motion with velocity w, and takes the value $\frac{1}{2} mw^2$. These specific facts have relevance only to the case when cross-sections move without change of shape; results corresponding to them will need to be derived for the extension of elongated-body theory to balistiform motion, which nevertheless will continue to be founded on basic principles which are essentially those set out in (a)-(h) above.

An important check on the validity of the conclusions derived from these principles (a)-(h) — together with the specific results (i)-(iii) for the case when cross-sections move without change of shape — was given when it was shown (*MB*, p. 78) that a small-perturbation expansion for the complete three-dimensional problem gives the same results. This is an expansion in powers of a parameter ε characterizing the magnitude of the fluid velocities, relative to the fish's forward speed, generated in the plane of a cross-section by the fish's motion through the fluid. In cases when this parameter is small, an expansion to second order in ε gives the same result for the force $\boldsymbol{F}$ as is derived by the principles (a)-(h). At the same time, the soundness of the physical arguments embodied in these principles encourages us, particularly after such a check in the small-parameter case, to use them also when relatively large cross-sectional motions are involved (*MB*, pp. 105-111). Similarly, it encourages us to apply to the case of balistiform swimming a set of principles which differ from those in (a)-(h) above in only one (quite minor) respect.

This one change is made only to principal (d). Here, the criteria governing our choice for the orientation of the posterior plane Π no longer constrain it to remain perpendicular to the backbone.

The essential need, in fact, is for this plane Π to be the plane of the fluid motions in the neighbourhood of the posterior cross- section, so that the fluid momentum vector $\boldsymbol{M}$ lies in the plane Π . An important difference, however, between swimming by body undulations and balis-

tiform swimming is that in the latter case the fluid momentum vector $\boldsymbol{M}$ is found (see §4) to be in a direction quite different from the direction perpendicular to the backbone specified by (i) above. Now, Π must be taken as a vertical plane through this momentum vector $\boldsymbol{M}$; and, accordingly, is directed obliquely to the backbone.

For balistiform swimming, furthermore, the correct posterior location for the plane Π is no longer the tip of the caudal fin, because the caudal fin is held rigid and plays no part in the propulsion of the fish. Instead, Π needs to be placed at the posterior end of the propulsive apparatus provided by the dorsal and anal fins.

The reasons for this relocation are two-fold. First, it does not interfere with the momentum-budget method (*b*) for finding the propulsive force $\boldsymbol{F}$ because the whole of this force is exerted within the volume $\mathcal{V}$ anterior to the plane Π; accordingly, it can be related to the rate of change of fluid momentum within $\mathcal{V}$. But the location proposed for Π is not only acceptable: it is also essential in order that the shed vorticity [see (e) above] is found entirely behind Π.

For balistiform swimming, in which the body and caudal fin remain rigid, the mathematical expression of the momentum budget (*b*) takes the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^L \boldsymbol{M}\,\mathrm{d}a = -\boldsymbol{F}+\boldsymbol{P}-U(\boldsymbol{M})_{a=0}, \tag{1}$$

where U is the forward velocity of the fish. Here, just as in the elongated-body theory for swimming by body undulations, the coordinate a signifies, for a cross-section of the fish anterior to the plane Π, its distance along the backbone[†] from the location of Π ; that is, from the posterior end of the fish's propulsive apparatus, taken here as $a = 0$. The position $a = L$ can be taken as the anterior end of the region of undulating fin movements, since this is the only region where the momentum $\boldsymbol{M}$ per unit length defined by (g) above is non-zero.

Then the left-hand side of (1) represents the rate of change of the fluid momentum in $\mathcal{V}$ (which as just explained takes the value $\boldsymbol{M}$ per unit length of backbone between $a = 0$ and $a = L$). The right-hand side, as in (b) above, equates this rate of change to the force $(-\boldsymbol{F})$ with which the fish acts on the fluid, plus the pressure force $\boldsymbol{P}$ acting across the posterior

[†]Needless to say, the fact that the backbone is here straight makes the definition of a simpler than in the theory of swimming by body undulations.

plane Π, minus the rate of transport of momentum out of $\mathscr{V}$ across Π; since, from (c), there is none across the horizontal plane boundaries.

Here, the rate of transport of momentum out of $\mathscr{V}$ across Π takes the form of the product UM (the distance forward travelled by Π in unit time multiplied by the momentum per unit distance). Exactly as in the theory for swimming by body undulations it is only the translational motion of the plane Π that produces momentum transport across it. Its rotational motion about a vertical axis produces by contrast no momentum transport (compare *MB*, p.109) because the distribution of momentum in the plane Π is an even function of distance from that vertical axis.

Analysis of how to calculate the quantities M and P which enter into (1) for the propulsive force is given in §4 (see also Parts 2 and 3). It is applied to calculate the mean propulsive force over a single cycle of the fin undulations. Just as for cases of body undulations, the expression on the left-hand side of (1) has zero mean since, when the time-derivative is integrated with respect to time over a cycle, we obtain the difference between the values of the cyclically varying quantity

$$\int_0^L M \, da \tag{2}$$

at the beginning and end of the complete cycle — values which are necessarily the same. Accordingly, the mean propulsive force is

$$\bar{F} = -\overline{U(M)}_{a=0} + \bar{P}, \tag{3}$$

where the bars signify means and we shall find that both terms on the right-hand side make a positive contribution to the propulsive force.

This result (3) for the mean propulsive force has its direct analogy for cases of body undulations. In those cases, however, the form (1) of the equation specifying the force F also gives an important indication of the magnitude of those fluctuating sideforces which cause an overall lateral oscillation of the entire fish body (*MB*, pp. 85-92); and this in turn produces an augmentation of that body drag which the propulsive force has to overcome.

By contrast, balistiform swimming involving several wavelengths of fin undulation may be expected to produce rather little fluctuating sideforce, particularly since the integral (2) whose time-derivative appears in (1) is expected to remain small. Essentially, this is because in an

integration over several wavelengths the contributions from the crests and troughs of the waves can be expected to cancel out.

This conclusion is important because it suggests why balistiform swimming is able to avoid those lateral oscillations which can augment body drag. Indeed, one of the two main advantages of this mode of swimming may (as suggested in §1) be that it keeps body drag down to the value associated with rigid forward motion. The other main advantage is indicated by calculations of (3) for mean propulsive force given in the next section.

4. A simplified model of balistiform locomotion

The extension of elongated-body theory given in §3 is applied first of all to a simplified model of balistiform locomotion. This is a model in which the thickness of the fish's (essentially rigid) body is taken small compared with its depth; the body, in short, is treated simply as a flat plate. Later, however, the theoretical conclusions for this simplified model are largely generalized (see §5 and, especially, Part 3) to bodies of elliptical cross-section with arbitrary ratios of thickness to depth.

In both models — whether with an elliptical or a 'thin-plate' body cross-section — the median fins (dorsal and anal) are, we emphasize, taken as thin but capable of a specific mode of flexure. This reflects the fact that the skeleton of either fin is composed of rays, rather flexibly joined together, and each provided with muscle fibres that can cause the ray to make a lateral movement in the approximate form of a rotation about its base.

For a particular cross-section, on the present simplified model,Figure 4 illustrates a sequence of four phases in the (synchronized) motion of the dorsal and anal fins. Here, the symmetrical configurations (*a*) and (*c*) represent the two phases when fin movements are at maximum velocities, so that the associated momentum $\boldsymbol{M}$ per unit length is expected to reach its peak. By contrast, in the extreme configurations (*b*) and (*d*), the fins are instantaneously at rest so that any irrotational motion around the cross-section ceases (being dependent only on the instantaneous movement of the boundary) and therefore $\boldsymbol{M}=0$.

Our objective in this section, then, is to calculate $\boldsymbol{M}$ in configuration (*a*) and in its mirror-image configuration (*c*), configurations representing those twin peaks in the magnitude of $\boldsymbol{M}$ during a cycle that appear between the phases (*b*) and (*d*) when $\boldsymbol{M}$ vanishes[†]. First of all we perform this calculation for a very precisely two-dimensional flow associated with

[†]Similarly, the calculations in Part 3 for bodies of elliptical cross-section are carried out for analogous symmetrical configurations of median fins.

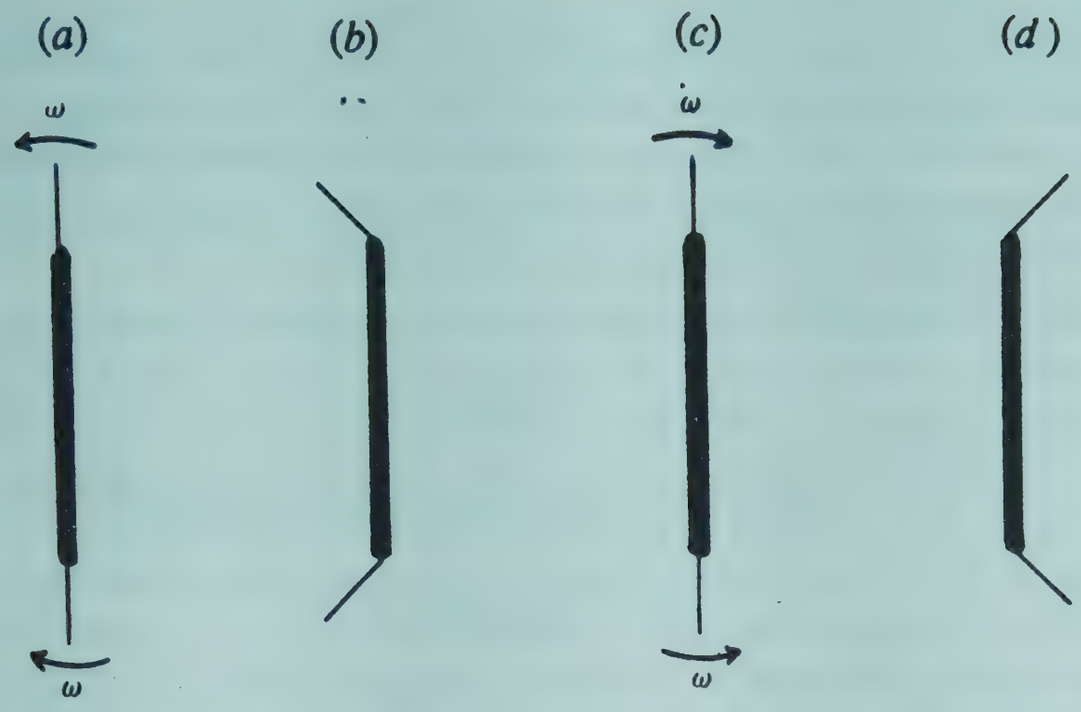


Fig. 4. Successive configurations of a fish cross-section in the simplified model of balistiform locomotion: (*a*) symmetrical configuration with maximum angular velocity ω of fin movement; (*b*) extreme configuration with fins instantaneously stationary; (*c*) mirror-image symmetrical configuration with reversed angular velocity ω; (*d*) mirror-image extreme configuration with stationary fins.

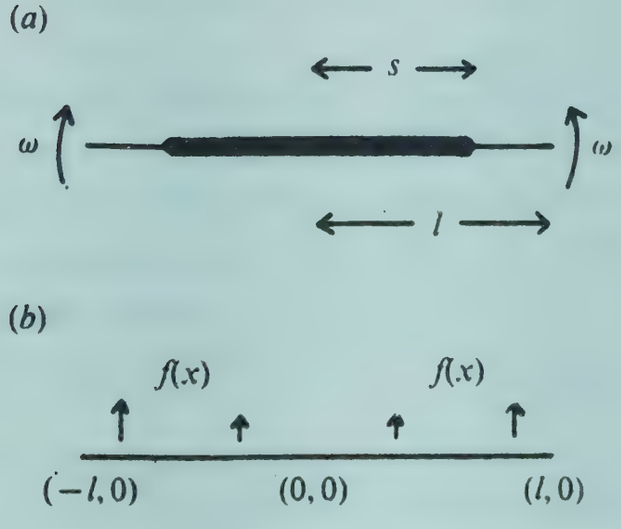


Fig. 5. (*a*) The very precisely two-dimensional problem associated with either configuration (*a*) or configuration (*c*) of Figure 4. (*b*) Generalization of the above to the case of a flexible flat plate of length $2l$ with velocity $f(x)$ normal to itself.

configuration (*a*). This represents the seriously unrealistic case when the motions of neighbouring cross-sections are all in phase (so that each median fin flaps rigidly back and forth like a door about its hinge); later, we extend it to a truly three-dimensional type of balistiform motion (of undular form) passing along each fin, with a phase lag in the motions of posterior sections behind those of anterior sections.

The precisely two-dimensional problem of calculating the momentum $\boldsymbol{M}$ associated with either configuration (*a*) or configuration (*c*) is illustrated in Figure 5 (turned through 90° in each case because this makes the fluiddynamic problem more familiar). Figure 5 also generalizes the problem to one in which the motion of a flexible flat plate $-l < x < l$ perpendicular to itself has velolcity $f(x)$. The case discussed above, with the inflexible body occupying the internal region $-s < x < s$ and the fins rotating rigidly with angular velocity ω about their bases $x = \pm s$, would be represented by taking the function $f(x)$ as

$$f(x) = 0 \quad (|x| < s), \quad \omega(|x| - s) \quad (s < |x| < l). \tag{4}$$

Mathematically, the problem reduces to solving the two-dimensional Laplace equation for the velocity potential ϕ outside the plate given that

$$\partial\phi/\partial y = f(x) \quad (-l < x < l). \tag{5}$$

The required momentum $\boldsymbol{M}$ per unit length is directed along the y-axis, with magnitude

$$M = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} \rho(\partial\phi/\partial y)\, \mathrm{d}y \right] \mathrm{d}x = \rho \oint \phi \, \mathrm{d}x, \tag{6}$$

where the order of integration (with respect to y first, and then to x) reflects the choice of $\boldsymbol{M}$ as the momentum between two parallel planes $|x| = X$ with X large, and where the integral on the right-hand side of (6) represents an integral around the surface of the plate in the positive (anticlockwise) sense.

The solution is well known when $f(x)$ is a simple constant, say $f(x) = 1$ (so that the entire plate moves with unit velocity). With the expression ϕ_1 used to mean the form of the potential ϕ in this classical case, its value on the surface of the plate is given as

$$\phi_1 = \mp (l^2 - x^2)^{\frac{1}{2}}, \tag{7}$$

with the $\mp$ sign standing for – on the upper surface and + on the lower surface of the plate. The integral (6) then gives M as $\rho \pi l^2$, the added mass for uniform motion of the plate.

This solution ϕ_1 for $f(x) = 1$ has one particular property — the fact that, by (5), $\partial \phi_1 / \partial y = 1$ on the plate itself — which allows us to calculate the momentum (6) for general $f(x)$. Evidently, we can rewrite (6) as

$$M = \rho \oint \phi (\partial \phi_1 / \partial y)\, dx \tag{8}$$

and then use Green's theorem to rewrite it as

$$M = \rho \oint \phi_1 (\partial \phi / \partial y)\, dx. \tag{9}$$

This, by (5) and (7), gives M as

$$M = 2\rho \int_{-l}^{l} (l^2 - x^2)^{\frac{1}{2}} f(x)\, dx. \tag{10}$$

since in the integral (6) around the plate the change in x is negative on the upper and positive on the lower surface.

The form of (10) already makes clear (for the present simplified model) the main conclusion of this paper: that a pair of moving fins, each of depth $l - s$, attached to a rigid body of depth $2s$ will generate considerably more momentum than would be derived by simply adding up the momentum that the motion of each fin on its own would produce. This is because in the integral(10)the size of the factor $(l-x)^{1/2}(l+x)^{1/2}$ on (say) the fin where $s < x < l$ is enhanced to take into account the total depth $2l$ of the body and fins combined - whereas the corresponding factor for the fin on its own would be only $(l-x)^{1/2}(x-s)^{1/2}$.

More precisely, the form (4) of the function $f(x)$ yields for the momentum (10) the expression

$$M = 2\rho\omega[\tfrac{1}{3}(l^2 - x^2)^{\frac{1}{2}}(2l^2 + s^2) - sl^2 \cos^{-1}(s/l)]; \tag{11}$$

and Figure 6 plots the ratio

$$\beta = M/M_0, \quad \text{where} \quad M_0 = \tfrac{1}{4}\pi\rho\omega(l-s)^3 \tag{12}$$

stands for the combined momentum that would be produced by the motion of each of the two fins on its own. Here, β can be described as a 'momentum enhancement factor', which is seen to lie between 3 and 4

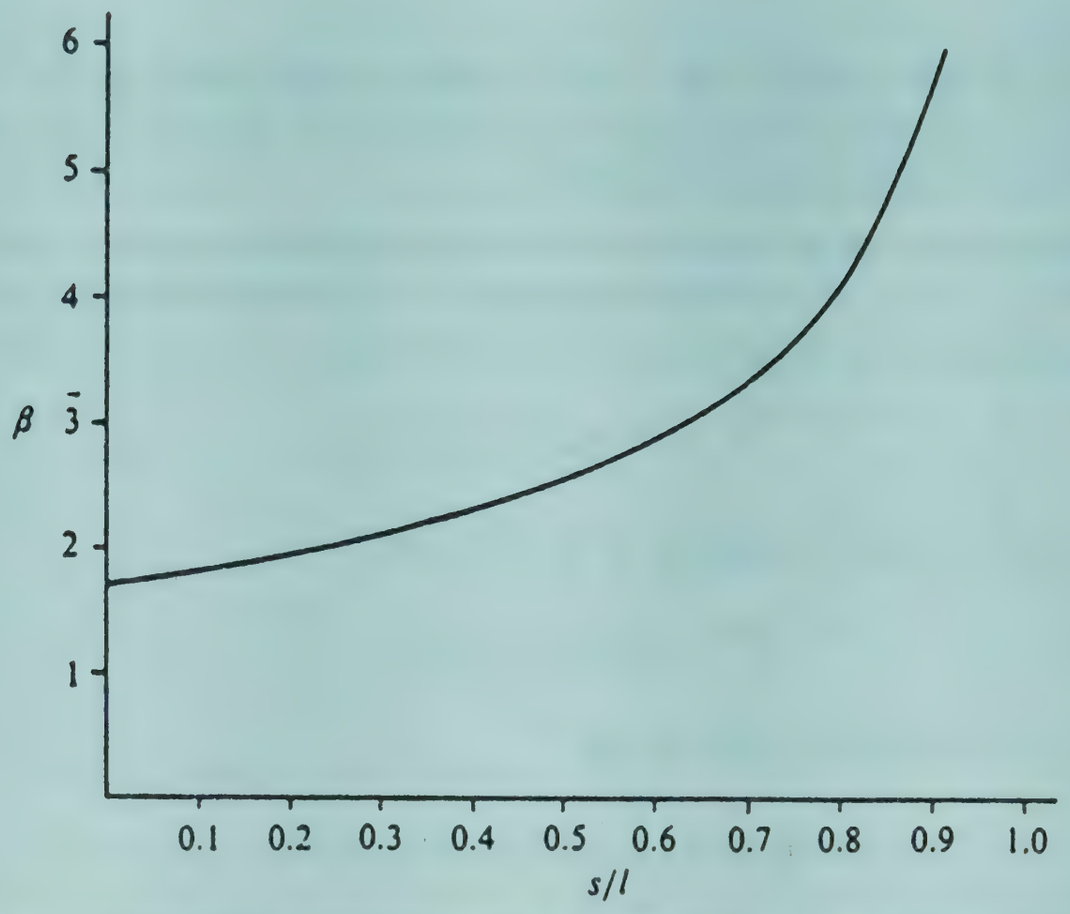


Fig. 6. The momentum enhancement factor β, as defined in (12), is plotted as a function of *s/l* for the simplified model shown in Figure 5.

for typical values of s / l between 0. 62 and 0. 78. This represents a very substantial enhancement.

On the other hand, all of the above calculations apply only in the case of 'very precisely two-dimensional flow'. Not only is this a seriously unrealistic case as noted earlier, but we have found also that it is a case where M is directed laterally (that is, along the y-axis) and therefore reverses its direction between configurations (*a*) and (*c*) — making, accordingly, no contribution to the mean quantity ($-\overline{UM}$) which appears in the expression (3) for propulsive force.

In the real three-dimensional balistiform motion, however, the undular character of the fin movements implies a phase lag in the motions of posterior sections behind those of anterior sections, and this twists the plane of the locally almost two-dimensional motion so that the momentum vector M is twisted always towards the posterior direction. Such a twist, as we shall see, makes the term ($-\overline{UM}$) contribute positively to the propulsive force.

In clarifying these matters, we again focus attention, at each fish cross-section, on configuration (*a*) — and therefore also, by implication,

on its mirror-image configuration (*c*). Figure 7 illustrates configuration (*a*) at a certain cross-section, alongside the associated configurations at posterior and anterior cross-sections with, respectively, phase lag and phase lead in their movements. An identical twist towards the posterior

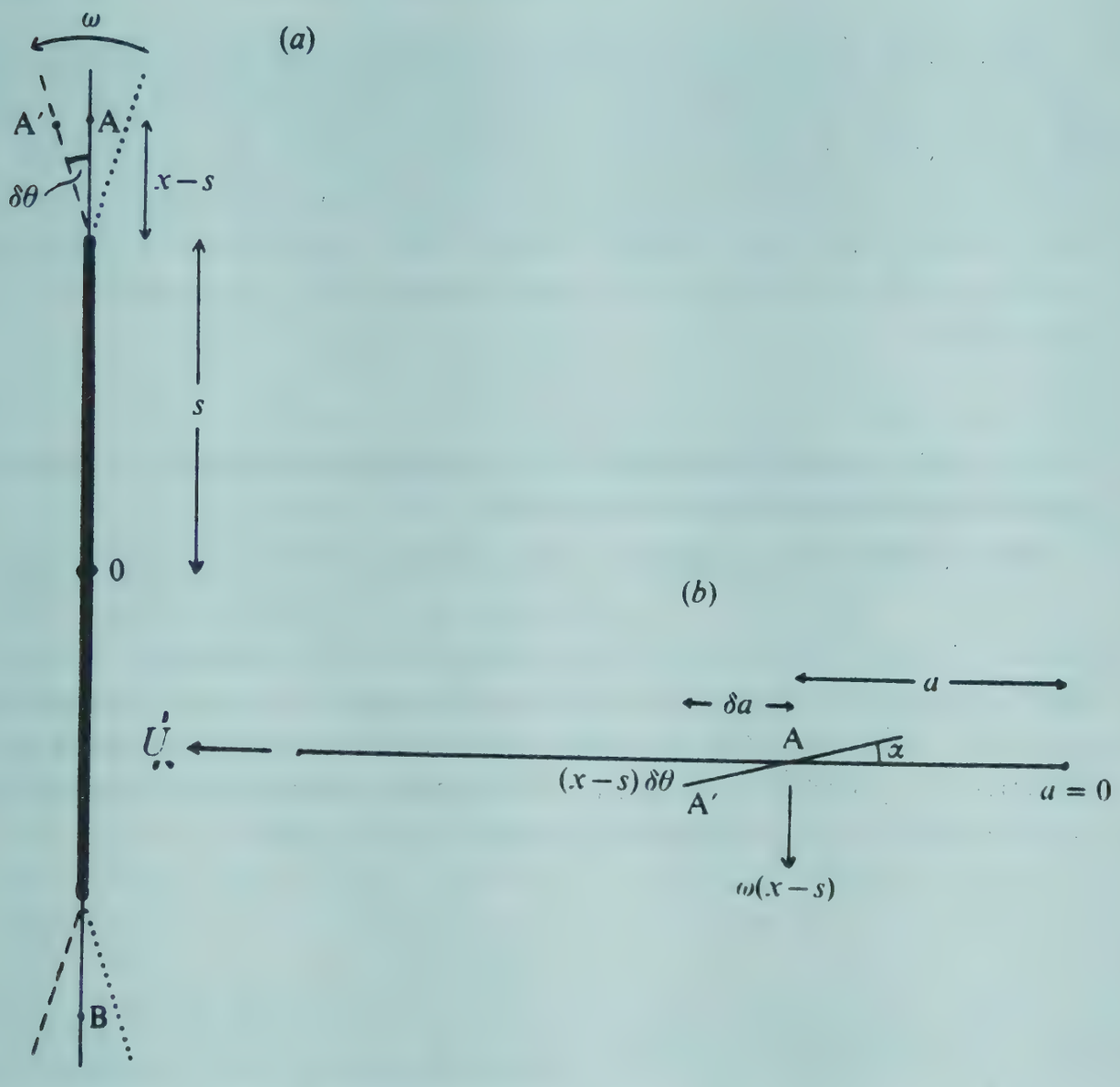


Fig. 7. (*a*) View of fins from posterior end (drawing 'in elevation'). _____, Fin positions at a distance *a* from posterior end, when the fins are in configuration (*a*) (see Figure 4). The points A and B are at a distance *x*-*s* from the fin base. -----, Fin positions with phase lead, found at a bigger distance $a+\delta a$ from the posterior end. Here the point A′ is at a distance *x*-*s* from the fin base. ....., Fin positions with phase lag, found at a smaller distance $a-\delta a$ from the posterior end. (*b*) View of fin section AA′ from above (drawing 'in plan'). Since [as (*a*) shows] the point A′ is displaced by a distance $(x\text{-}s)\delta\theta$ from the vertical plane through the backbone, where $\delta\theta$ is the (small) angle shown, the angular displacement α of AA′ satisfies the equation $\tan\alpha = [(x\text{-}s)\delta\theta]/\delta a$. The motions of AA′ are lateral with velocity $\omega(x\text{-}s)$ and forward with velocity *U*.

direction in the motions of both fins perpendicular to themselves is evident from this diagram.

We denote by κ the local twist of each fin, defined (see Figure 7) as

$$\kappa = \delta\theta/\delta a. \tag{13}$$

Here, with a as the distance of a certain section ahead of the posterior end of the fish's propulsive apparatus and $a + \delta a$ as the same distance for an anterior section, $\delta\theta$ represents the angle between the fin positions at the two sections when projected as in Figure 7 onto the same plane. (If, for example, the fin's undular motion took the form of a sinusoidal variation of θ with amplitude θ_0 and wavelength λ, then the value of κ would be

$$\kappa = 2\pi\theta_0/\lambda; \tag{14}$$

but the more general definition (13) allows for the possibility of the fin's motions not being precisely sinusoidal.)

We consider now, at a point A with coordinates $(x, 0)$ so that its distance from the fin base is $(x - s)$, the angle α of inclination of the fin (see Figure 7) to the fixed vertical plane through the fish's backbone. Because, at the section anterior to A by a distance δa the fin is turned through an angle $\delta\theta$, it follows that the point corresponding to A is displaced by a distance $(x - s)\,\delta\theta$; accordingly, the tangent of the angle of inclination α is the ratio of $(x - s)\,\delta\theta$ to δa, which by (13) gives

$$\tan\alpha = \kappa(x - s). \tag{15}$$

Clearly, the direction normal (i.e. perpendicular) to the fin at A is itself inclined at the same angle α to the lateral direction (which is the direction normal to the above mentioned vertical plane). Both at A and at the corresponding point B on the other fin — with coordinates $(-x, 0)$ — the normal instead of being in the lateral direction is twisted towards the posterior direction through this same angle α.

It follows that in configuration (a) the motion of each fin normal to itself at both A and B is directed not laterally but in another direction derived by angular displacement towards the posterior direction through an angle α. The magnitude of that motion (at A, for example) consists, as Figure 7 shows, of two terms: the resultant $\omega(x - s)\cos\alpha$ of the fin's lateral motion with velocity $\omega(x - s)$ minus the resultant $U \sin\alpha$ of the

fin's forward motion at the swimming velocity U of the fish. Equation (15) for $\tan\alpha$ allows us to express the resultant in a form

$$\omega(x-s)\cos\alpha - U\kappa(x-s)\cos\alpha = (1-V^{-1}U)\,\omega(x-s)\cos\alpha, \qquad (16)$$

where we have written $V = \omega / \kappa$.

We may recall that the multiplying factor $(1-V^{-1}U)$ occurs also in classical elongated-body theory (*MB*, p. 20) with V as the velocity of propagation of the propulsive undular motion along the fish's body. This is also the interpretation of V in (16), because an undular variation in θ propagating at speed V in the posterior direction (with a decreasing) takes the form

$$\theta = f(a+Vt), \quad \text{giving} \quad \partial\theta/\partial t = V\,\partial\theta/\partial a \quad \text{or} \quad \omega = V\kappa. \qquad (17)$$

We have found that the motion of each fin normal to itself occurs with velocity (16) and is directed not laterally but in another direction turned through an angle α towards the posterior direction. We must now ask what this implies for the approximate specification of the local fluid motion by a two-dimensional motion.

It is evident first of all that, if α were a constant instead of varying along the fin $s < x < l$ in accordance with (15), then the local approximately two-dimensional motion (produced by the motion of each fin normal to itself) would be not in the plane of the cross-section but in a plane derived from it by such a backward twist through angle α. This is because the fin movements normal to themselves would be in such a plane (while the fish's essentially rigid body does not move at all); accordingly, the consequent fluid motion and its momentum M would be in the same plane, and would therefore possess a posterior-pointing component (contributing positively, as already noted, to the term $(-\overline{UM})$ in the expression (3) for propulsive force).

Next, if we take into account the actual variation with x in the angle α through which the motions are twisted, we must ask in what plane we can then expect the momentum vector M, related (as far as possible) to some approximately two-dimensional fluid motion in that plane, to lie. We note first that the motion should be as nearly two-dimensional as possible if the plane is defined as one twisted backward through an angle α_{av} representing some sort of average of the angles α defined by (15). Furthermore, the momentum vector M should then be approximately the momentum per unit length associated with that two-dimensional motion,

since fin movements in the region (near the fin base with $\alpha < \alpha_{av}$) would push the fluid to one side of that plane while those in the region (near the fin tip) with $\alpha > \alpha_{av}$ would push it to the other side, and with a proper choice of α_{av} the associated components of fluid momentum normal to the plane should cancel out.

Indeed, this argument suggests what type of average should be used in calculating α_{av}; namely, a weighted average for $s < x < l$ with the velocity (16) as the weighting function. Figure 8 plots the angle α_{av} so calculated, as a function of the product $\kappa(l - s)$ of the twist (13) and the depth of a fin.

Although the three-dimensional effects discussed above suggest an angular displacement α_{av} (given by Figure 8) for the direction of the momentum M per unit length of fish, they do not give rise as regards its magnitude M to any prediction very greatly modified from that (given by Figure 6) emerging from the two- dimensional theory. The one important change is the incorporation of the $(1-V^{-1}U)$ factor from (16).

This equation, indeed, implies that, at a point (such as A in Figure 7) with coordinates $(x, 0)$, the fin's velocity component in the plane of the approximately two-dimensional motion can be written as

$$(1 - V^{-1}U)\,\omega(|x| - s)\cos\alpha_{av}. \tag{18}$$

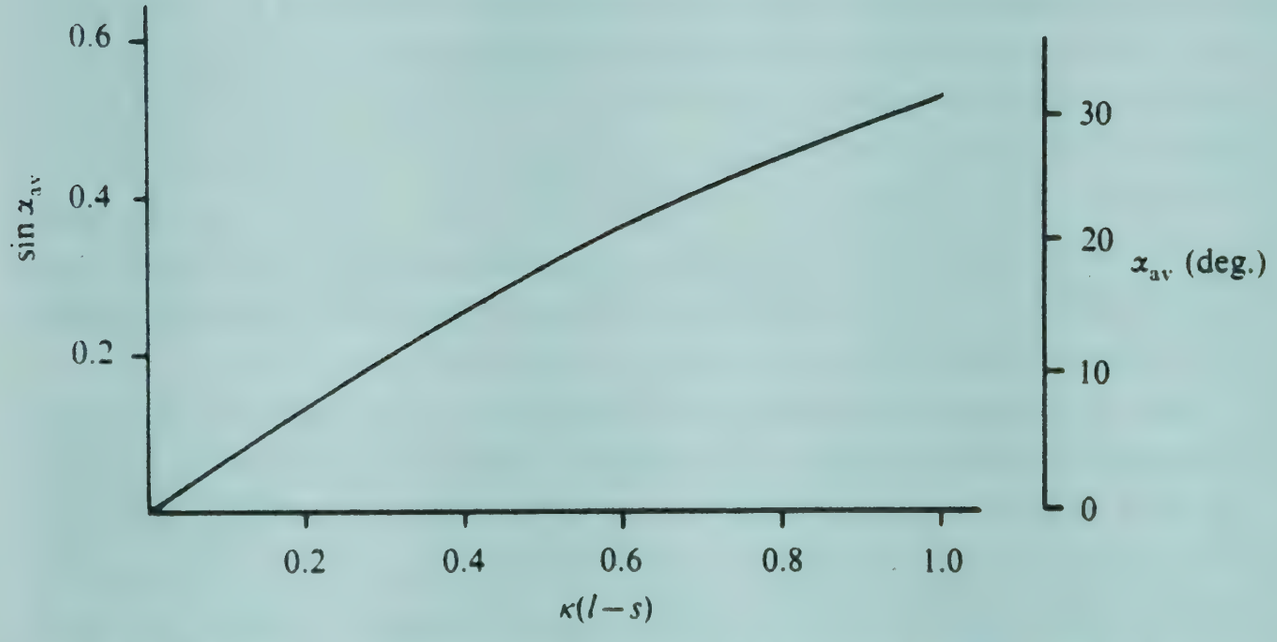


Fig. 8. The angular displacement α_{av} calculated as a weighted average [with the velocity (16) used as a weighting function] of the angle α defined by (15). The right-hand scale gives α_{av} in degrees, while the left-hand scale gives sin α_{av} [a factor appearing in the principal contribution (21) to the propulsive force].

We illustrate in Figure 9 a short length δa of fish on which the lines at angle α_{av} to the backbone show how the actual fin movements are represented as closely as possible by movements (18) giving rise to a locally two-dimensional motion. These movements represent a reduction by a factor $(1-V^{-1}U)\cos\alpha_{av}$ in the fin velocities (4) assumed in the earlier, strictly two- dimensional theory. Accordingly, the momentum per unit length of the fin itself, given by (12) as βM_0, is reduced to

$$\beta M_0(1-V^{-1}U)\cos\alpha_{av}. \tag{19}$$

However, the fin length shown in Figure 9 is $(\delta a)\sec\alpha_{av}$ and so the overall momentum is $\beta M_0(1-V^{-1}U)\delta a$, giving

$$M=\beta M_0(1-V^{-1}U) \tag{20}$$

as the magnitude of the momentum per unit length of fish.

In summary, then the momentum M in configuration (a) has magnitude M and is directed at an angle α_{av} to the lateral direction. As noted earlier, its strictly lateral component $M\cos\alpha_{av}$ is reversed in the mirror-image configuration (c), and therefore is not expected to contribute to the mean value $\overline{UM}$. On the other hand, the momentum component

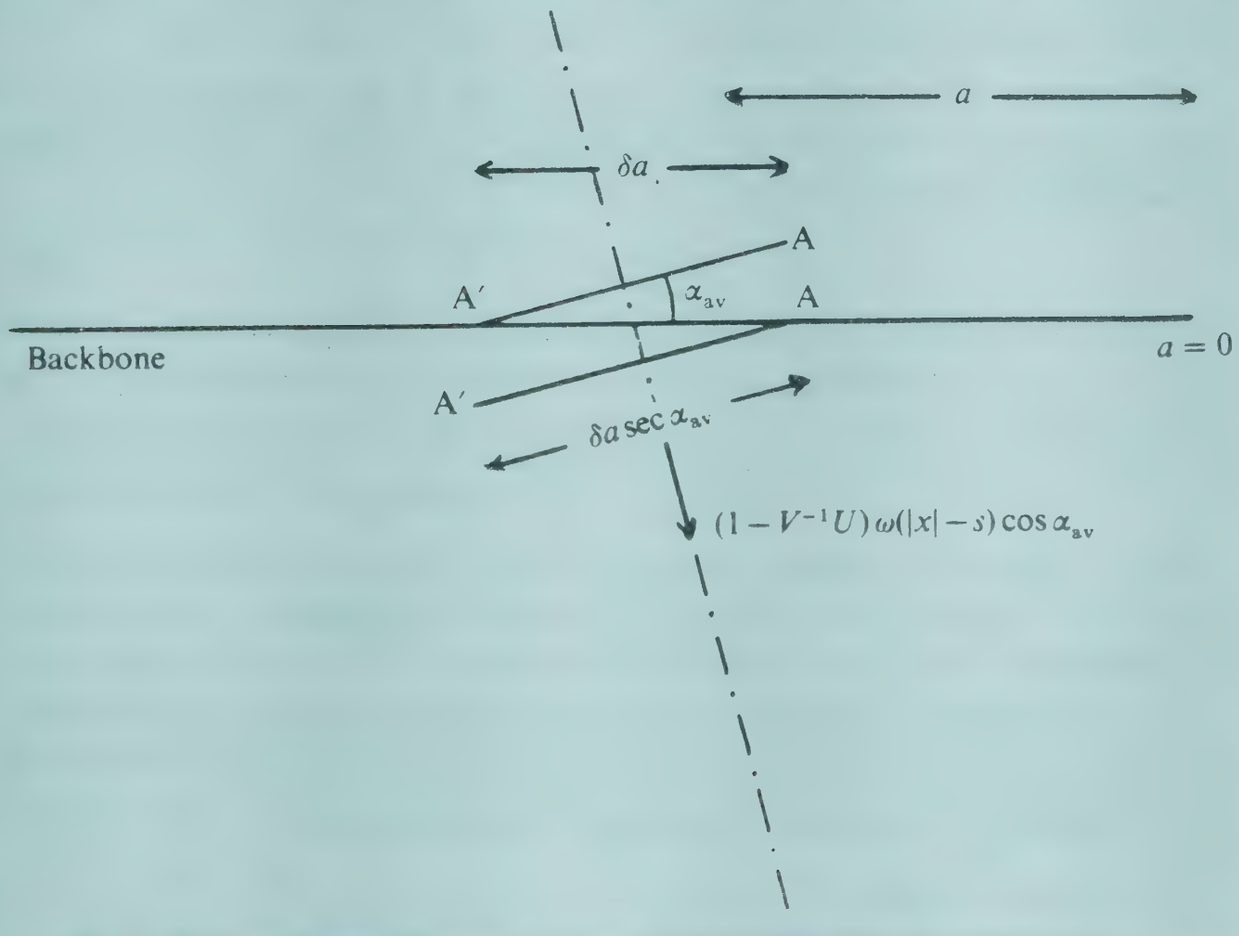


Fig. 9. View from above (drawing 'in plan') of the motion of the fin section AA′. . _.-._.-._, Plane of the associated approximately two-dimensional motion of fluid.

$M \sin \alpha_{av}$ in the posterior direction is unchanged in that mirror-image configuration (*c*), in between configurations (*b*) and (*d*) when both components are zero; and, for the corresponding component of UM, this represents a fluctuation at twice the fundamental frequency between $UM \sin \alpha_{av}$ and zero, with a best estimate of its mean value as $(1/2)UM \sin\alpha_{av}$ [where (20) gives the approximate value of M]. The corresponding term ($-\overline{UM}$) in the propulsive force (3) is, because of the minus sign, directed forwards with the same magnitude.

$$\tfrac{1}{2}U\beta M_0(1 - V^{-1}U)\sin\alpha_{av}. \tag{21}$$

This term (21) seems to be the leading contributor to propulsive force. It corresponds directly to the term $\overline{umw\,\partial z/\partial a}$ in large-amplitude elongated body theory [*MB*, equation(5) on p. 109], with U corresponding to u and $\sin \alpha_{av}$ to $\partial z / \partial a$; but the momentum enhancement factor β, plotted in Figure 6, increases considerably the importance of this term(21).

By contrast, the other term in the propulsive force(3) — namely, the mean of the pressure force $\boldsymbol{P}$ acting across the posterior plane — is recognised in large-amplitude elongated-body theory as of only minor importance. Furthermore, it contributes even less in the present theory because it turns out not to be subject to any such enhancement.

To make this clear, we note that in Part 2 the magnitude of $\boldsymbol{P}$ is calculated for a very precisely two-dimensional motion [with velocities as in (4) above)] in the form

$$P = \zeta P_0, \quad \text{where} \quad P_0 = \rho\omega^2(l-s)^4 \tag{22}$$

and the coefficient ζ takes values ranging (see Figure 10) from 0.53 to 0.62 instead of becoming specially enhanced as s/l approaches 1. The replacement of ω by $\omega(1 - V^{-1}U)\cos\alpha_{av}$ in (18) changes (22) into

$$P = \zeta(1 - V^{-1}U)^2 P_0 \cos^2\alpha_{av}, \tag{23}$$

where, just as in classical elongated-body theory, the reduction factor $(1 - V^{-1}U)$ appears in a squared form that contributes to the relative insignificance of P. The vector pressure force $\boldsymbol{P}$ is directed at right angles to the plane of the motions so that its forward component is $P \cos \alpha_{av}$ in configurations (*a*) and (*c*)— compared with zero values in configurations (*b*) and (*d*) — and the best estimate of its mean value is $\frac{1}{2} P \cos \alpha_{av}$, giving

$$\tfrac{1}{2}\zeta(1 - V^{-1}U)^2 P_0 \cos^3\alpha_{av} \tag{24}$$

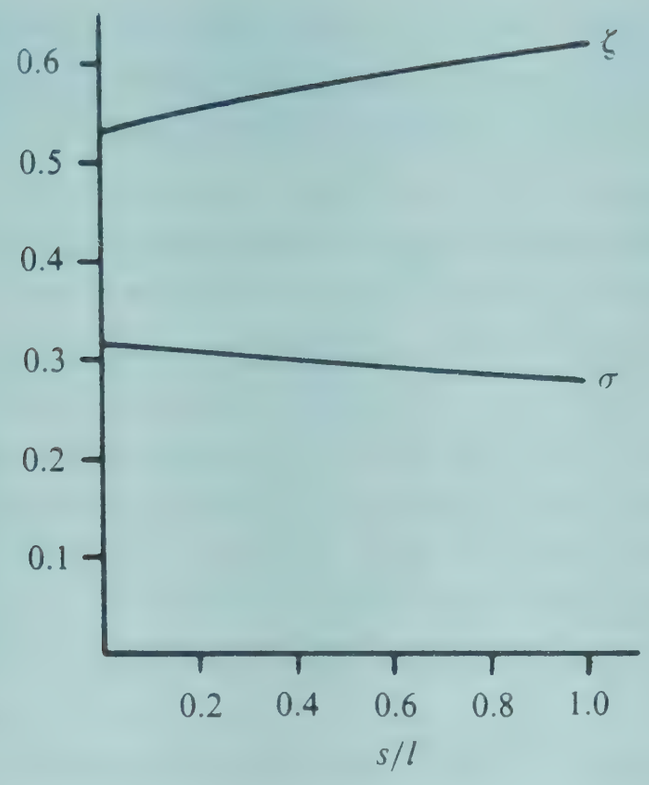


Fig. 10. Effective absence of enhancement (as s/l increases) in the pressure force P or in the kinetic energy per unit length E. The graphs, taken from Figure 1 of Part 2, show the variation of $\zeta = P/P_0$ and of $\sigma = E/P_0$ as a function of s/l for the simplified model illustrated in Figure 5, where $P_0 = \rho\omega^2(l-s)^4$.

as the contribution of $\overline{P}$ in (3) to the propulsive force.

We can confirm the unimportance of this contribution (24) by writing out its ratio to the leading term(21) as

$$\frac{\frac{1}{2}\zeta(1-V^{-1}U)^2\rho\omega^2(l-s)^4\cos^3\alpha_{av}}{\frac{1}{2}U\beta[\frac{1}{4}\pi\rho\omega(l-s)^3](1-V^{-1}U)\sin\alpha_{av}} \tag{25}$$

[using (22) and (12) for P_0 and M_0) and substituting $\omega = V\kappa$ from (16). Then the ratio (25) becomes

$$\left(\frac{4\zeta}{\pi\beta}\right)(VU^{-1}-1)\left[\frac{\kappa(l-s)}{\sin\alpha_{av}}\right]\cos^3\alpha_{av}\,; \tag{26}$$

and, with ζ between 0.53 and 0.62 and β in the region of 3 to 4, the quantity in the first bracket is small; and that in the second is also expected to be small, while the product of the remaining factors (see Figure 6) is close to 1 or a bit less.

To sum up, the propulsive force (3) is dominated by the term (21) representing the rate of shedding of the backward-pointing component of

fluid momentum at the posterior end of the fish's propulsive apparatus. This momentum is enhanced by a substantial factor β which represents the effect of the fin movements being attached to a deep, essentially rigid, body.

Finally, the rate of shedding of kinetic energy at the posterior end of the fish's propulsive apparatus is not so enhanced; which, of course, has favourable implications for the efficiency of balistiform propulsion. In Part 2, indeed, for a very precisely two-dimensional motion [with velocities as in (4) above], the kinetic energy per unit length is calculated in the form $E = \sigma P_0$, where the coefficient σ takes values which vary only slightly (see Figure 10), from 0.32 to 0.28. It follows that, in the approximately two-dimensional motion illustrated in Figure 9, with ω replaced by $\omega(1-V^{-1}U)\cos\alpha_{av}$, the kinetic energy in a length δa of fish becomes

$$[\sigma(1-V^{-1}U)^2 P_0 \cos^2\alpha_{av}](\delta a)\sec\alpha_{av}. \qquad (27)$$

Accordingly, the rate at which fin movements generate 'unproductive' energy which is shed in the vortex wake is

$$U\sigma(1-V^{-1}U)^2 P_0 \cos\alpha_{av} \qquad (28)$$

in configurations (*a*) and (*c*); or [see (21) above] half as much on a time-averaged basis.

When we compare this rate of generation of 'unproductive' energy with the rate of generation of 'useful' work by only the main term (21) in the propulsive force, we obtain once more a ratio as in (25), but with ζ replaced by the somewhat smaller coefficient σ and (in partial compensation for this reduction) $\cos^3\alpha_{av}$ replaced by $\cos\alpha_{av}$. The resulting ratio is again small (a bit smaller than before) and for two main reasons:

(i) As in classical elongated-body theory, the factor $(1-V^{-1}U)$ appears to the first power in the rate of generation of useful work but squared in the rate of generation of unproductive energy; while

(ii) The ratio is still further reduced by the presence in the denominator of the momentum enhancement factor β .

That factor represents additionally, then, an enhancement of propulsive efficiency.

5. Generalizations and conclusions

Analysis of balistiform swimming in §4 by means of a simplified model led to some clear conclusions: that propulsive force is dominated

by the rate of shedding of backward momentum, and that a substantial momentum enhancement factor β both magnifies this force and improves propulsive efficiency. We must now ask whether the model's principal geometrical simplification (characterization of the lateral thickness of the body as negligibly small) may have exaggerated the momentum enhancement effect.

The answer to this question is obtained in Part 3, §3, where fish bodies with realistically elliptical cross-sections of axis ratio t/s (ratio of the body thickness $2t$ to the body depth $2s$) are analysed and the corresponding momentum enhancement factor β derived as a function of t/s and of the solidity ratio s/l (ratio of the body depth $2s$ to the total cross-sectional depth $2l$ including fins as well as body). Part 3 illustrates the sectional geometry at the top of its Figure 6 and sets out the results in its Figures 7 and 8. We see from Figure 7 of Part 3 that calculations for the simplified model (with t/s essentially zero) do exaggerate the momentum enhancement effect but only to a modest extent. Figure 8 of Part 3 shows indeed that, as t/s increases from 0 to as high a value as 1 (case of a circular body cross-section), the momentum enhancement factor drops from 2.9 to 2.6 when $s/l = 0.6$, from 3.4 to 2.8 when $s/l = 0.7$ and from 4.2 to 3.0 when $s/l = 0.8$. We recall that balistiform locomotion is exhibited mainly in fishes with values of t/s somewhat intermediate[†] between 0 and 1. The above calculations caused us to comment in §1 above that momentum enhancement for balistiform locomotion is by a factor of 'around 3 or a little more'.

The propulsive force in balistiform swimming should be dominated, then, by the momentum-shedding term

$$\tfrac{1}{2}U\beta M_0(1 - V^{-1}U)\sin\alpha_{av} \qquad (29)$$

(see 21)) with the above revised values of β. We have not considered it worthwhile to try to recalculate the additional term $\bar{P}$ which, from being insubstantial in classical elongated-body theory, is made negligible (§4) as a result of momentum enhancement.

We emphasize that our conclusions apply equally to both of the two main modes of balistiform locomotion, designated in Figure 1 as the

[†]At the same time, because the plane in which the fluid motions are most closely two-dimensional is inclined (see figure 9) at an angle α_{av} to the fish cross-section, the effective value of t/s for the body's elliptical section by such a plane is increased by a factor sec α_{av}, and it is for this increased value of t/s that the appropriate value of β needs to be read off from the curves in Part 3.

undulatory and oscillatory modes. The force depends on the value of (29) at the posterior section $a = 0$ of the fish's propulsive apparatus; and this value is well defined in terms of ω, the angular velocity of fin-ray movement and κ, the twist (13) of each fin, as that section passes through the symmetrical configuration (a). For a truly undulatory motion, the ratio $V = \omega / \kappa$ can be interpreted (see 17)) as the propagation speed of the wave; but the ratio remains well-defined also in the oscillatory mode, when its interpretation is really only as a sort of 'equivalent wave speed'. (A similar distinction exists for modes of locomotion by body flexure (*MB* , p. 20), where a certain quantity V represents a true wave speed in the anguilliform mode but only an equivalent wave speed in the carangiform mode.)

Momentum enhancement is present, then, to much the same extent for both modes of balistiform locomotion; and we are forced to ask, therefore, what advantages, if any, may have accompanied the evolution of those more complicated neuromuscular control mechanisms that are required for the undulatory mode. The main benefit, already noted towards the end of §1 and at the end of §3, stems from the virtual elimination of fluctuating sideforces - allowing the essentially rigid body to move forward without those oscillating motions of sideslip or yaw which are known (Blake 1983*a* , pp. 98-101) to increase body drag by a factor of 3 to 4. Provisionally, then, we may expect the above-noted enhancement of thrust by a factor of about 3 or a bit more to coexist with reduction in body drag by a comparable factor, which, with body drag proportional to U^2, might enhance the swimming velocity U by (again) a similar factor.

Here we should pause and consider whether any other serious inaccuracies in our model (besides the geometrical one which has been comprehensively addressed in Part 3) may limit the value of the above conclusions. In this context their most obvious source of inaccuracy lies in the approach to determining $\boldsymbol{M}$ through its calculation for a locally two-dimensional flow that approximates as well as possible the real three-dimensional flow.[†] It is in the undulatory case, particularly if the wavelength λ of undulations is not very large, that the approximate

[†]By contrast, the neglect of any vorticity shed at fin tips is likely to reduce only slightly the accuracy of the irrotational-flow calculation of $\boldsymbol{M}$ for an unsteady fluid motion like this, where the rather small amplitude of oscillation of fins relative to the water limits greatly the displacement of any shed vorticity.

representation of solutions of the three-dimensional Laplace equation by solutions of the two-dimensional equation

$$\partial^2\phi/\partial x^2 + \partial^2\phi/\partial y^2 = 0 \qquad (30)$$

may involve substantial errors. These indeed are cases where a different two-dimensional equation

$$\partial^2\phi/\partial x^2 + \partial^2\phi/\partial y^2 - k^2\phi = 0 \quad (\text{where } k = 2\pi/\lambda) \qquad (31)$$

might be thought to represent the three-dimensional Laplace equation more accurately.

Although we have not embarked on any calculations using (31), preferring at present to estimate momentum enhancement by a direct comparison of the results of applying elongated-body theory to different cases, we can see from a knowledge of the fundamental solution $K_0(kr)$ of (31) in Bessel-function form where $(x^2 + y^2)^{1/2} = r$) how that equation's use might affect the calculation of momentum. Basically, this solution behaves like the corresponding solution of (30) at short range but falls off far more rapidly when kr is large. Accordingly, the momentum produced by given fin movements will be spread over a gradually reduced area as k increases. Momentum enhancement, in short, may be expected to be reduced as the wavelength of an undulatory motion is itself reduced. This consideration, particularly in relation to studies of the evolution of the undulatory form of balistiform locomotion, may impose a countervailing disadvantage on any further reductions in wavelength after the initial advantage of drag limitation through reductions in body oscillations has been reaped.†

We should not leave the subject of balistiform locomotion without referring to its use by certain groups of flexible-bodied fishes as an alternative low-cost means of progression when they do not need the higher speeds available with anguilliform body undulations. The thin bodies of the flat-fishes (Heterosomata) make the model of §4 rather accurate (even though the whole geometry is turned through 90°), and the conclusions in that section regarding efficiency confirm the low energy cost of propulsion of such fishes by simple undulations of the marginal fins. Analogous advantages are reaped by many eels (Apodes)

† As Part 1 goes to press we add this footnote confirming that calculations based on equation (31), which have now been made, do indeed demonstrate such a 'countervailing disadvantage'. These calculations and their biological implications are to be set out in Part 4 of the present series.

which are able to hold their body straight while wave motions pass along their median fins either in the posterior (or for backward motion) in the anterior direction; although in their case, of course, the model of Part 3 is appropriate with t / s around 1.

Finally, we touch briefly on the biofluiddynamics of gymnotiform locomotion in fishes possessing a long ventral fin but no corresponding dorsal fin; that is, a mode of locomotion in which the fish's body is held rigid and undulations pass along the ventral fin in the posterior direction. Our analysis follows the same form as in §4, with the omission of any dorsal-fin effects; but, needless to say, a new determination of the momentum enhancement factor is needed.

Part 3 gives this in its §2 for fish bodies with elliptical cross-sections of axis ratio t / s, and obtains the momentum enhancement factor β as a function of t / s and of the solidity ratio $2s / (l + s)$, where the latter is now the ratio of the body depth $2s$ to a total cross-sectional depth $l + s$ including the depth $l - s$ of just a single fin. Part 3 illustrates the sectional geometry in its figure 2 and sets out the results in its figures 4 and 5. The momentum enhancement factor is here defined as

$$\beta = M/M_1, \quad \text{where } M_1 = \tfrac{1}{8}\pi\rho\omega(l-s)^3 \tag{32}$$

is the momentum that would be associated with the motion of just the single fin on its own [thus, comparison with (12) gives $M_1 = \frac{1}{2} M_0$] The value of β for typical solidity ratios around 0. 7 is in the region of 2.4, representing a considerable momentum enhancement but not quite so much as in the balistiform mode. The estimate (21) of propulsive force in the balistiform mode has to be replaced by

$$\tfrac{1}{2}U\beta M_1(1 - V^{-1}U)\sin\alpha_{av} \tag{33}$$

in the gymnotiform mode.

Very much as before, we see the advantages of gymnotiform locomotion as residing partly in an enhancement of thrust (without a corresponding enhancement of 'unproductive' energy generation) and partly in a reduction of drag, the latter reduction being evident once again from the argument given at the end of §3. These are substantial advantages, leading to a reduced energy cost of transport, and are available, of course, whether or not a fish possesses electric organs whose use may be facilitated in a mode of locomotion with the body held rigid.

REFERENCES

Batchelor G.K. 1967 *An Introduction to Fluid Dynamics*. Cambridge University Press.

Blake, R.W. 1981 Mechanics of ostraciiform locomotion. *Can. J. Zool.* **59**, 1067-1071.

Blake, R.W. 1983a *Fish Locomotion*. Cambridge University Press.

Blake, R.W. 1983b Swimming in the electric eels and knifefishes. *Can. J. Zool.* **61**, 1432-1441.

Blake, R.W. 1983c Median and paired fin propulsion. In *Fish Biomechanics* (ed. P.W. Webb & D. Weihs). Praeger.

Gray, J. 1968 *Animal Locomotion*. Weidenfield & Nicolson.

Lighthill, J. 1975 *Mathematical Biofluiddynamics*. Philadelphia: Society for Industrial Applied Mathematics (Referred to as *MB*).

Lightill, J. 1986 *An Informal Introduction to Theoretical Fluid Mechanics*. Oxford University Press.

Lighthill, J. 1990a Biofluiddynamics of balistiform and gymnotiform locomotion. Part 2. The pressure distribution arising in two-dimensional irrotational flow from a general symmetrical motion of a flat plate normal to itself. *J. Fluid Mech.* **213**, 1-10.

Lighthill, J. 1990b Biofluiddynamics of balistiform and gymnotiform locomotion, Part 3. Momentum enhancement in the presence of a body of elliptic cross-section. *J. Fluid Mech.* **213**, 11-20.

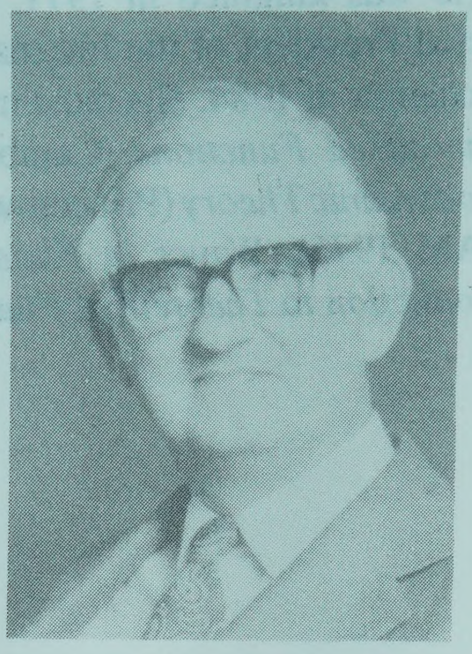
Lissmann, H.W. 1958 On the function and evolution of electric organs in fish. *J. Exp. Biol.* **35**, 156-191.

Marshall, N.B. 1965 *The Life of Fishes*, Weidenfeld & Nicolson.

Marshall, T.C. 1964 *Fishes of the Great Barrier Reef*. Angus & Robertson.

Webb, P.W. & Weihs, D. (eds.) 1983 *Fish Biomechanics*. Praeger.

Accession No. 488
4 FEB 1994
REGIONAL CENTRE, LUDHIANA



SIR JAMES LIGHTHILL

MICHAEL JAMES LIGHTHILL was born on 23rd January 1924. After studying mathematics at Cambridge (1941-1943) he entered the Aerodynamics Division of the National Physical Laboratory where he pursued original work in both subsonic and supersonic aerodynamics.

He became a Fellow of Trinity College, Cambridge in 1945 and a year later, Senior Lecturer in the University of Manchester and subsequently Beyer Professor of Applied Mathematics, in 1950. Here, he continued to be involved in extensive aerodynamic and hydrodynamic research, while also initiating the science of aerodynamic sound generation (with the object of allowing jet engines to become more powerful without becoming noisier).

In 1959 he assumed the Directorship of the Royal Aircraft Establishment based at Farnborough, where he promoted the development of Concorde, of VTOL aircraft, and of space technology. In 1964 he became Royal Society Research Professor at Imperial College and soon after, Physical Secretary of the Royal Society, while turning his research attention to marine science and to biofluiddynamics. He occupied the Lucasian Chair of Applied Mathematics at Cambridge (once held by Sir Isaac Newton) from 1969 to 1979; when he was appointed Provost of University College London. He retired from the Provostship in 1989 and is now an Honorary Research Fellow in the Department of Mathematics at University College London.

He was knighted in 1971, becoming Sir James Lighthill, and was elected President of the International Union of Theoretical and Applied Mechanics in 1984. His most important books are: *Fourier Analysis and Generalised Functions* (Cambridge 1957), *Higher Approximations in Aerodynamic Theory* (Princetown 1960), *Mathematical Biofluiddynamics* (SIAM 1975), *Waves in Fluids* (Cambridge 1978), and *An Informal Introduction to Theoretical Fluid Mechanics* (Oxford 1986).